Современный
Гуманитарный
Университет

Дистанционное образование

0664.03.01;1

Рабочий учебник

Фамилия, имя, отчество _____

Факультет _____

Номер контракта _____

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

ЮНИТА 3

ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Москва 2001

Экземпл.

Разработано И.Б. Чернышевой, канд. техн. наук

Рекомендовано Министерством общего и профессионального образования Российской Федерации в качестве учебного пособия для студентов высших учебных заведений

КУРС: ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Юнита 1. Введение в теорию вероятностей.
Юнита 2. Многомерные распределения и предельные теоремы.
Юнита 3. Основные понятия математической статистики.
Юнита 4. Прикладная статистика.

ЮНИТА 3

Рассматриваются основные понятия математической статистики: выборочный метод, построение точечных оценок для параметров распределения, доверительные интервалы, проверка статистических гипотез. В приложении дан список формул по теории вероятностей и математической статистике.

Для студентов Современного Гуманитарного Университета

Юнита соответствует профессиональной образовательной программе № 1

(С) СОВРЕМЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ, 2001

ОГЛАВЛЕНИЕ

ДИДАКТИЧЕСКИЙ ПЛАН 5
ЛИТЕРАТУРА 6
ПЕРЕЧЕНЬ УМЕНИЙ 7
ТЕМАТИЧЕСКИЙ ОБЗОР 12
Введение 12
1. Выборочный метод 14
1.1. Выборка 14
1.2. Методы отбора. Репрезентативность выборки. Выборка повторная и бесповторная 15
1.3. Вариационный ряд. Группировка. Табличное представление выборки 18
1.4. Графическое представление выборки. Полигон, гистограмма, кумулята 23
1.5. Числовые характеристики выборки 25
1.6. Связь между статистическим распределением выборки и изучаемым распределением вероятностей 28
1.6.1. Полигон и многоугольник распределения 29
1.6.2. Гистограмма и плотность вероятности 30
1.6.3. Кумулята, эмпирическая и теоретическая функции распределения 32
2. Построение точечных оценок для параметров распределения 34
2.1. Распределения вероятностей, зависящие от параметра 34
2.2. Метод моментов 34
2.3. Вычисление эмпирических моментов 41
2.4. Свойства точечных оценок 42
2.5. Понятие надежности оценки 43
2.6. Распределение выборочного среднего 44
2.7. Связь между точностью и надежностью оценки 45
3. Интервальные оценки для параметров 50
3.1. Понятие доверительного интервала 50
3.2. Доверительный интервал для среднего в случае, когда среднеквадратическое отклонение σ теоретического распределения известно 51
3.3. Доверительный интервал для среднего в случае, когда среднеквадратическое отклонение σ теоретического распределения неизвестно 52
3.4. Оценка требуемого объема выборки 54
3.5. Доверительный интервал для вероятности успеха в схеме Бернулли 55

3.6. Односторонние доверительные интервалы 57
4. Проверка статистических гипотез 61
4.1. Проверка гипотезы о равенстве среднего числовому значению 61
4.2. Сравнение двух генеральных средних 64
4.3. Ошибки первого и второго рода. Мощность критерия . . 70
4.4. Оценка требуемого числа испытаний при проверке гипотезы 72
5. Методы, основанные на свойствах функции правдоподобия 74
5.1. Функция правдоподобия 74
5.2. Метод максимального правдоподобия для получения оценки неизвестного параметра Θ 74
5.3. Функции правдоподобия в задачах проверки гипотез . 76
5.4. Проверка гипотез методом последовательного анализа 78
ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ 81
ТРЕНИНГ УМЕНИЙ 86
ФАЙЛ МАТЕРИАЛОВ 101
ГЛОССАРИЙ*

* Глоссарий расположен в середине учебного пособия и предназначен для самостоятельного заучивания новых понятий.

ДИДАКТИЧЕСКИЙ ПЛАН

Предмет математической статистики. Метод сплошных наблюдений и выборочный метод. Понятие выборки. Табличное и графическое представление выборки – полигон и гистограмма.

Выборочное среднее и его закон распределения, выборочная дисперсия. Точечные оценки и их свойства: состоятельность, несмещенность. Идея метода моментов. Числовые характеристики случайных величин и их вероятностный смысл. Некоторые распределения: биномиальное, пуассоновское, равномерное, показательное, нормальное.

Интервальное оценивание. Распределение Стьюдента. Доверительные интервалы для генерального среднего нормальной величины с известной и неизвестной дисперсией. Доверительный интервал для вероятности успеха в схеме Бернулли. Односторонние доверительные интервалы.

Проверка статистических гипотез. Уровень значимости, ошибки первого и второго рода. Методы проверки статистических гипотез: с использованием доверительных интервалов, методом Неймана-Пирсона, методом последовательного анализа.

ЛИТЕРАТУРА

Базовая

1. Гмурман В.Е. Теория вероятностей и математическая статистика. М., 2000.

2. Кремер Н.Ш. Теория вероятностей и математическая статистика. М., 2000.

Дополнительная

3. Калинина В.Н., Панкин В.Ф. Математическая статистика. М., 1998.

4. Янко Ярослав. Математико-статистические таблицы. М., 1961.

Задачники

5. Гмурман В.Е., Руководство к решению задач по теории вероятностей и математической статистике. М., 1997.

6. Булдык Г.М., Мацкевич И.П., Свирид Г.П. Сборник задач и упражнений по высшей математике (теория вероятностей и математическая статистика). Минск, 1996.

Примечание. Знаком (*) отмечены работы, на основе которых составлен тематический обзор.

ПЕРЕЧЕНЬ УМЕНИЙ

№ п/п	Умение	Алгоритмы
1.	Построение по выборке таблицы распределения, полигона и гистограммы	1. Упорядочить заданные значения по возрастанию, сосчитать их количество. 2. Если надо, сгруппировать значения; сосчитать число значений, попавших в интервалы разбиения; вычислить эмпирические частоты; составить таблицу эмпирического распределения. 3. По таблице эмпирического распределения нарисовать гистограмму и полигон, найти медиану.
2.	Вычисление точечных оценок параметров распределения по выборке	1. Выписать заданные значения, объем выборки и нужную формулу для получения точечной оценки. *Указание к шагу:* Вычисление точечной оценки для среднего производится по формуле: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ – для выборки, заданной вариационным рядом, и $\bar{x} = \frac{1}{n}\sum_{j=1}^{k} x_j m_j = \sum_{j=1}^{k} x_j \frac{m_j}{n} = \sum_{j=1}^{k} x_j \tilde{p}_j$ – для выборки, заданной таблицей. Вычисление смещенной точечной оценки для дисперсии производится по формуле: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ – для выборки, заданной вариационным рядом, и по формуле: $S^2 = \frac{1}{n}\sum_{j=1}^{k} x_j^2 m_j - \bar{x}^2 = \sum_{j=1}^{k} x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^{k}(x_j - \bar{x})^2 \frac{m_j}{n}$ – для выборки, заданной таблицей. Вычисление несмещенной точечной оценки для дисперсии производится по формуле: $s^2 = \frac{n}{n-1} S^2$ $\sigma = \sqrt{S^2}$ – оценка для среднеквадратического отклонения. 2. Сосчитать значение оценки.

№ п/п	Умение	Алгоритмы
3.	Вычисление доверительных интервалов для среднего	1. Сосчитать выборочное среднее, выборочное среднеквадратическое отклонение (если не известно истинное), выписать нужную формулу доверительного интервала. *Указание к шагу (пользоваться таблицами файла материалов):* Формула доверительного интервала для математического ожидания μ нормального распределения с уровнем доверия β для случая, когда известно среднеквадратическое отклонение распределения σ: $$\bar{x} - k_\beta \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + k_\beta \frac{\sigma}{\sqrt{n}},$$ где k_β находится из табл. 5 нормального распределения по заданному уровню доверия β. Если среднеквадратическое отклонение неизвестно, вместо σ надо использовать его эмпирическую оценку S с заменой в формуле доверительного интервала n на $n-1$ и вместо значений k_β нормального распределения использовать значения $t_{n-1,\beta}$ распределения Стьюдента с n степенями свободы, содержащиеся в Таблице 6: $$\bar{x} - t_{n-1,\beta} \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_{n-1,\beta} \frac{S}{\sqrt{n-1}},$$ где $t_{n-1,\beta}$ находится с помощью табл. 6. Ту же формулу можно выписать через несмещенную оценку s $$\bar{x} - t_{n-1,\beta} \frac{s}{\sqrt{n}} < \mu < \bar{x} + t_{n-1,\beta} \frac{s}{\sqrt{n}}.$$ 2. Пользуясь табл. 5 или 6 вычислить границы требуемого в задании интервала, выписать полученный доверительный интервал.

№ п/п	Умение	Алгоритмы
4.	Вычисление доверительного интервала для вероятности p наступления события A с помощью таблиц нормального распределения	1. Вычислить оценку $\tilde{p} = \frac{m}{n}$ для p. 2. Найти доверительный интервал для p. *Указание к шагу:* Основная формула – следствие интегральной теоремы Муавра-Лапласа, из которой выводятся любые соотношения между эмпирической частотой, генеральной частотой, n и вероятностью β: $P\left\{\left\|\frac{m-np}{\sqrt{npq}}\right\| < k_\beta\right\} = P\left\{\ |m-np| < k_\beta \cdot \sqrt{npq}\right\} =$ $= P\left\{\left\|\frac{m}{n} - p\right\| < k_\beta\sqrt{\frac{pq}{n}}\right\} = P\left\{\left\|\frac{m}{n} - p\right\| < k_\beta\sqrt{\frac{p(1-p)}{n}}\right\} \cong$ $\cong \Phi(k_\beta) = \beta$. Доверительный интервал для p ищется по формуле: $\tilde{p} - k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ — выборка с повтором, $\tilde{p} - k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\sqrt{1-\frac{m}{N}} \le p \le \tilde{p} + k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\sqrt{1-\frac{m}{N}}$ – выборка без повтора, где $k_\beta = 1{,}96$ для уровня доверия 95% и и $k_\beta = 3$ для уровня доверия 99,7% (табл. 5). 3. В случае, когда требуется, проверить гипотезу, сформулировать вывод из эксперимента, провести вычисления с доверительным интервалом и т.д.
5.	Проверка статистических гипотез	1. Выписать из условия задачи данные о выборке. Сосчитать оценки для среднего и дисперсии. 2. Сформулировать проверяемую гипотезу в вероятностных терминах. Выписать формулу статистики, вычисляемой по выборке. Выписать число степеней свободы N для распределения статистики. Подставить в формулу статистики данные выборки.

№ п/п	Умение	Алгоритмы
		Указание к шагу: Проверка гипотезы производится на заданном уровне значимости α. Изучаются два варианта: 1) Выборка из одной совокупности, ее параметр (среднее) сравнивается с известным значением. То есть проверяется гипотеза H_0: $\mu = \mu_0$. Альтернативными гипотезами H_1 могут быть: a) $\mu \neq \mu_0$ b) $\mu > \mu_0$ c) $\mu < \mu_0$ По выборке вычисляется значение статистики $$T = \frac{\overline{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$ Число степеней свободы $N = n - 1$, где n – объем выборки. 2) Сравниваются параметры двух генеральных совокупностей. Из обеих делаются выборки, проверяется гипотеза H_0: $\mu_x = \mu_y$. Альтернативными гипотезами H_1 могут быть: a) $\mu_x \neq \mu_y$ b) $\mu_x > \mu_y$ c) $\mu_x < \mu_y$ По выборке вычисляется значение статистики: $$T = \frac{\overline{x} - \overline{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\overline{x} - \overline{y}}{\sqrt{nS_x^2 + mS_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}},$$ где $$s^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i - \overline{x})^2 + \sum_{j=1}^{m}(y_i - \overline{y})^2\right] =$$ $$= \frac{1}{n+m-2}(nS_x^2 + mS_y^2)$$ Число степеней свободы $N = n + m - 2$, если n – объем выборки из X, а m – объем выборки из Y.

№ п/п	Умение	Алгоритмы
		3. Выписать критическую область и с помощью таблиц найти границы критической области для статистики, с помощью которой будет проверяться гипотеза. *Указание к шагу:* a) Гипотеза H_0: $\mu_1 = \mu_2$, альтернативная H_1: $\mu_1 \neq \mu_2$. Критическая область для проверки гипотезы: $$\lvert T \rvert > t_{N;\alpha}$$ ($t_{N;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в верхней строке). Если вычисленное значение не попало внутрь интервала $[-t_{N;\alpha}, t_{N;\alpha}]$, то гипотеза H_0 не проходит при уровне значимости α. b) Гипотеза H_0: $\mu_1 = \mu_2$, альтернативная H_1: $\mu_1 > \mu_2$. Критическая область для проверки гипотезы: $$T > t_{N;\alpha}$$ ($t_{N;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке). Если вычисленное значение попало внутрь интервала $[t_{N;\alpha}, \infty]$, то гипотеза H_0 не проходит при уровне значимости α. c) Гипотеза H_0: $\mu_1 = \mu_2$, альтернативная H_1: $\mu_1 < \mu_2$. Критическая область для проверки гипотезы: $$T < -t_{N,\alpha}$$ ($t_{N,\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке) Если вычисленное значение попало внутрь интервала $[-\infty, -t_{N,\alpha}]$, то гипотеза H_0 не проходит при уровне значимости α. 4. Проверить, попало или нет в критическую область значение статистики. Сформулировать вывод, требуемый в задаче.

ВВЕДЕНИЕ

В основе всех научных знаний лежит наблюдение. Для обнаружения общей закономерности, которой подчиняется явление, необходимо многократно его наблюдать в одинаковых условиях. Например, начальник цеха изучает вопрос о проценте брака для изделий, обработанных на некотором станке. Обследуется 100, 1000 изделий. Сколько должно быть проведено наблюдений? Как обработать результаты наблюдений и сделать обоснованные практические выводы? Или такой пример. Исследователя интересует зависимость урожайности определенной культуры от количества внесенных удобрений и качества обработки почвы. Для выяснения этой зависимости собраны сведения об урожайности, количестве внесенных удобрений и качеству обработки по достаточно большому числу одинаковых участков. Как, используя эти сведения, оценить зависимость урожайности от количества удобрений и условий обработки почвы? В обоих приведенных примерах, а также и во многих других явлениях, можно отметить, что, несмотря на постоянство условий испытания, результат опыта неоднозначен. Детали обрабатываются вроде бы одинаково, однако одни из них удовлетворяют требованиям приемки, другие – нет. Урожай, выращенный на "одинаковых" участках, – различен, и так далее. Предвидеть результат каждого конкретного опыта нельзя. Однако если систематизировать результаты измерений, то можно увидеть в их изменении некоторую закономерность, которая называется статистической устойчивостью. И хотя предвидеть результат каждого конкретного опыта нельзя, оказывается можно предвидеть в среднем результат серии измерений. Изучением закономерностей случайных явлений, а мы привели примеры именно случайных явлений, занимается теория вероятностей. Она строит математические модели случайных явлений, основываясь на формально логических рассуждениях. Например, в результате опыта может произойти одно из n событий, ни одно из которых не имеет перед другими никакого преимущества (бросается монета или игральный кубик). Логически рассуждаем – их вероятности одинаковы. Построенная теоретически модель позволяет вычислять вероятности наступления сложных случайных событий, представляющих интерес для практики. Например, какова вероятность, что, играя в монетку, сделав 100 попыток, я не про-

* Жирным шрифтом выделены новые понятия, которые необходимо усвоить. Знание этих понятий будет проверяться при тестировании.

играю весь свой капитал? Если говорить коротко, теория вероятностей позволяет находить вероятности "сложных" событий через построенные теоретически вероятности "простых" событий. Математическая же статистика оперирует результатами наблюдений над случайными явлениями для того, чтобы оценить их вероятности, либо с помощью серии опытов осуществляет проверку предположений относительно этих вероятностей. В самом общем виде то, чем занимается математическая статистика, можно описать так.

Математическая статистика – раздел математики, изучающий методы сбора, систематизации и обработки наблюдений с целью выявления статистических закономерностей.

Математическая статистика, опираясь на вероятностные модели, в свою очередь, влияет на развитие теории вероятностей. Математическая статистика и теория вероятностей – две неразрывно связанные науки.

У истоков статистической науки стояли две школы – немецкая описательная и английская школа политических арифметиков. Школа политических арифметиков зародилась в XVII веке. Именно представителями школы политических арифметиков была осознана необходимость учета в статистических исследованиях требований закона больших чисел, поскольку закономерность может проявиться лишь при достаточно большом объеме анализируемой совокупности. В XIX веке бельгийский статистик Кетле положил основание учению о средних величинах. Своим дальнейшим развитием математическая статистика обязана П.Л.Чебышеву, А.А.Маркову, А.М.Ляпунову, а также К.Гауссу, Ф.Гальтону, К.Пирсону и др. Существенный вклад в математическую статистику был сделан в XX веке советскими математиками (В.И.Романовский, Е.Е.Слуцкий, А.Н.Колмогоров, Н.В.Смирнов), а также английскими (Стьюдент, Р.Фишер, Э.Пирсон) и американскими (Ю.Нейман, А.Вальд) учеными.

1. ВЫБОРОЧНЫЙ МЕТОД

1.1. Выборка

Итак, проводится обследование совокупности объектов относительно некоторого качественного или количественного признака. Например, если имеется партия деталей, то качественным признаком может служить стандартность детали, а количественным – размер детали.

Иногда проводят сплошное обследование, то есть обследуют каждый объект совокупности. **Метод сплошных наблюдений** – метод статистического обследования, при котором производится измерение *всех* элементов совокупности. Но если число объектов очень велико или если обследование объекта требует больших затрат или приводит к его уничтожению, то проводят несплошное, выборочное, обследование. **Выборочный метод** – метод статистического обследования, при котором из совокупности выбирают ограниченное число объектов и их подвергают изучению. Он применяется тогда, когда количество объектов велико или сплошное обследование невозможно в силу того, что обследование может привести к уничтожению объекта (например, чтобы узнать качество консервов, банку надо вскрыть), т.е. когда не хотят проводить полное обследование объекта. Примером сплошного наблюдения является изучение успеваемости студентов администрацией вуза, перепись населения, охватывающая все население страны. Выборочными наблюдениями являются, например, социологические исследования, охватывающие часть населения.

Вся подлежащая изучению совокупность объектов (наблюдений) называется **генеральной совокупностью.** Та часть объектов, которая отобрана для непосредственного изучения из генеральной совокупности, называется **выборочной совокупностью** или **выборкой.**

Число объектов N генеральной совокупности и число объектов n выборочной совокупности называют **объемом** соответственно **генеральной и выборочной совокупности.**

Пример 1.1. Из партии товара, содержащей 10000 деталей, отобрано для обследования 100 деталей. Объем генеральной совокупности N = 10000, а объем выборки n = 100.

Естественно предполагать, что объем генеральной совокупности гораздо больше, чем объем выборки (N>>n).

Если объем генеральной совокупности достаточно велик, то иногда в целях упрощения вычислений или для облегчения теоретических выводов допускают, что генеральная совокупность состоит из бесчисленного множества объектов. Такое допущение оправдывается тем, что

увеличение генеральной совокупности, если ее объем достаточно велик, практически не сказывается на результатах обработки данных выборки. Таким образом, генеральная совокупность может иметь как конечный, так и бесконечный объем. Примером бесконечной совокупности может служить гипотетическая совокупность всех деталей, производимых заводом. В математической статистике понятие генеральной совокупности трактуется, вообще говоря, как совокупность всех мыслимых наблюдений, которые могли бы быть произведены при данном реальном комплексе условий. В математической модели, с которой оперирует математическая статистика, понятие генеральной совокупности в определенном смысле аналогично понятию случайной величины (вероятностному пространству, закону распределения вероятностей). Это не только имеющиеся в наличности объекты, но и все гипотетически возможные объекты, которые могли бы функционировать в том же комплексе условий. И выборка производится из "распределения вероятностей".

1.2. Методы отбора. Репрезентативность выборки. Выборка повторная и бесповторная

Выборочный метод исследования является единственно возможным в случае бесконечной генеральной совокупности или в случае, когда исследование связано с уничтожением наблюдаемых объектов. Кроме того, он позволяет существенно экономить затраты ресурсов. Недостатком его является появление ошибок исследования, (их называют ***ошибками репрезентативности***), которые *связаны с тем, что изучается только часть объекта*. Математическая статистика дает рекомендации, как организовать исследование, чтобы свести эти ошибки к минимуму, и дает методику оценки этих ошибок.

Чтобы по данным выборки иметь возможность судить о генеральной совокупности, выборка должна быть отобрана так, чтобы она давала правильное представление о генеральной совокупности.

Пример 1.2. Для проверки качества продукции отобрана партия втулок, изготовленная случайно выбранным рабочим. Но в цехе по производству втулок работают квалифицированные токари и начинающие. Ясно, что если эти втулки изготовлены квалифицированным токарем, то представление о качестве продукции, выпускаемой всем цехом, будет "завышенным", а если изучать втулки, изготовленные начинающим токарем, то "заниженным".

Для того, чтобы выборка давала представление о генеральной совокупности, необходимо, чтобы соблюдался принцип равной возможности всем элементам генеральной совокупности быть отобранными в

выборку. В приведенном примере в выборку попали втулки, изготовленные только одним рабочим, т.е. эти втулки при отборе имели преимущество и указанный принцип соблюден не был.

Выборка называется **репрезентативной (представительной** – от *англ.* representative), если она достаточно хорошо воспроизводит генеральную совокупность, т.е. это выборка, которая производится так, что все объекты генеральной совокупности имеют одинаковую вероятность попасть в выборку.

Обеспечить это условие можно различными средствами. Например, отбор можно производить просто на основе таблиц случайных чисел. Таких таблиц сейчас издано много; разработаны программы для ЭВМ – генераторы случайных чисел. Если изучается объект, состоящий из многих разнородных частей, например, мнение избирателей, надо позаботиться о том, чтобы в выборке в соответствующей пропорции были представлены все части системы. В ней должны быть представлены горожане и сельские жители, молодежь и пенсионеры, военные, рабочие, интеллигенция и т. д. из всех частей страны и в той же пропорции, что и во всей стране.

К чему может привести несоблюдение этого правила показывают многочисленные случаи несбывшихся предвыборных прогнозов. Например, в 1936 году перед президентскими выборами в США журнал "Literary Digest" провел опрос 10 миллионов избирателей и предсказал, что Франклин Рузвельт проиграет выборы. Фамилии избирателей были взяты из телефонных книг. Но в 30-е годы во время депрессии люди, имевшие телефон, не представляли всех избирателей США, выборка оказалась не репрезентативной и прогноз не оправдался. На телевидении вошло в моду проводить экспресс-опросы во время передачи – желающие сообщить свое мнение могут позвонить в студию и ответить на вопрос "да", "нет" или "не знаю". Такая форма опроса не дает репрезентативной выборки. Примером организации репрезентативного опроса может служить, в частности, метод отбора, который был применен в Англии при проведении обследования рациона питания среднего англичанина. Выборка извлекалась методом трехступенчатого отбора. На первом этапе было отобрано 50 избирательных округов. Затем из них было отобрано некоторое количество избирательных участков. На третьем – некоторое количество семей внутри этих участков. На каждом этапе отбор был строго случайным.

Существуют специальные приёмы отбора, обеспечивающие репрезентативность выборки.

Опишем простейшую схему получения репрезентативной выборки из конечной, не очень большой генеральной совокупности.

Все объекты генеральной совокупности нумеруют, номера записывают на отдельные карточки, карточки перемешивают и выбирают одну наудачу. Объект, номер которого совпал с номером на карточке, считается попавшим в выборку. Операцию повторяют до тех пор, пока не наберется нужный объем выборки. При этом, если случайно отобранная карточка возвращается обратно в общую совокупность и, следовательно, раз отобранный в выборку объект может быть отобран повторно, то имеет место **выборка повторная,** или **выборка с возвратом**, а если отобранная карточка и, следовательно, отобранный в выборку объект назад не возвращается, то осуществляется **выборка бесповторная** или **выборка без возврата.** Вместо перемешивания карточек, можно использовать таблицы случайных чисел. Их можно найти в большинстве книг по статистике. Следует заметить, что если в выборке с возвратом испытания независимы, то в выборке без возврата испытания уже зависимы. Для демонстрации разницы между этими схемами рассмотрим простейший пример.

Пример 1.3. В урне n белых и m черных шаров. Наугад вынимаем два шара. A_1 – событие, состоящее в том, что первый шар белый, A_2 – второй шар тоже белый.

Выборка с возвратом:

$$P(A_1) = P(A_2) = n/(n+m),\ P(A_2/A_1) = P(A_2) = n/(n+m).$$

Выборка без возврата:

$$P(A_1) = n/(n+m).$$

Для того, чтобы найти вероятность $P(A_2)$, определим событие B_1 – первый вынутый шар – черный и воспользуемся формулой полной вероятности:

$$P(A_2) = P(A_1)P(A_2/A_1) + P(B_1)P(A_2/B_1) =$$

$$= \frac{n}{n+m}\cdot\frac{n-1}{n+m-1} + \frac{m}{n+m}\cdot\frac{n}{n+m-1} = \frac{n}{n+m} = P(A_1).$$

Заметим, что

$$P(A_2/A_1) = \frac{n-1}{n+m-1} \neq P(A_2).$$

Таким образом, для выборки с возвратом вероятность во втором испытании вытащить белый шар такая же, как и в первом испытании: условная вероятность совпадает с безусловной, следовательно, испытания независимы. Для выборки без возврата вероятность во втором

испытании вытащить белый шар такая же, как и в первом испытании, но независимости испытаний уже нет. Легко видеть, что если n и m велики, то зависимость испытаний является слабой. Если объем генеральной совокупности достаточно велик, а выборка составляет лишь незначительную часть этой совокупности, то различие между повторной и бесповторной выборками стирается.

В дальнейшем будем предполагать, что требование репрезентативности выборки выполнено, испытания независимы и будем обсуждать только вопросы обработки выборочных данных.

1.3. Вариационный ряд. Группировка. Табличное представление выборки

Пусть из генеральной совокупности извлечена выборка объёмом n. Случайный выбор элемента рассматривается как независимое наблюдение над величиной ξ, имеющей некоторое распределение вероятностей. Если те значения y_1, $y_2 ... y_n$, которые приняла случайная величина ξ в n наблюдениях, записать не в порядке получения, а в порядке возрастания, то получим упорядоченную выборку $x_1, x_2, ... x_n$, называемую **вариационным рядом**. Наблюдаемые значения x_i называются **вариантами.**

Выборка и вариационный ряд несут одну и ту же информацию, но с вариационным рядом легче работать в силу его упорядоченности.

Расстояние $x_{max} - x_{min}$ между крайними членами вариационного ряда называется **размахом вариационного ряда**.

Если изучается дискретная случайная величина, то при достаточно большом объеме выборки в выборке будут повторяющиеся значения.

Для каждого полученного значения можно подсчитать, сколько раз оно встретилось в ряде наблюдений. Эти числа называются **частотой варианта,** или его **весом.** В дальнейшем частоту варианта x_i мы будем обозначать через m_i, где i – индекс варианта.

Данные наблюдений, среди которых много повторяющихся. удобно изобразить не в виде ряда, а в виде таблицы (табл. 1.1).

Таблица 1.1

Значения x_i	x_1	x_2	...	x_k
Частоты m_i	m_1	m_2	...	m_k

Пример 1.4. На телефонной станции проводились наблюдения над числом X неправильных соединений в минуту. Наблюдения в течение часа дали следующие результаты: 3; 1; 3; 1; 4; 2; 2; 4; 0; 3; 0; 2; 2; 0; 2; 1; ... 1; 1; 5. Расположив эти числа в порядке неубывания, получим следующий ряд: 0; 0; 0; 0; 0; 0; 0; 0; 1; 1; 1; ... 5; 5; 7. Значения 0; 1; 2; ..., 7, принятые случайной величиной в процессе наблюдений, являются вариантами.

Число в мин x_i	0	1	2	3	4	5	7	
Частоты m_i	8	17	16	10	6	2	1	$\Sigma = 60$

Назовем **относительной (эмпирической) частотой** значения x_i отношение m_i/n, где m_i – число повторения значения x_i (его частота) в выборке объема n. Относительные частоты – характеристика более универсальная, чем просто частоты, так как позволяет сравнивать выборки разного объема.

Построим по выборке таблицу из двух строк, в верхней строке которой указаны в порядке возрастания наблюдаемые значения x_i, а в нижней – соответствующие им относительные частоты.

Эта таблица называется **таблицей статистического распределения выборки** (табл. 1.2).

Таблица 1.2

Значения x_i	x_1	x_2	...	x_k
Относительные частоты m_i/n	m_1/n	m_2/n	...	m_k/n

Для примера 1.4 таблица статистического распределения выборки имеет вид:

Число неправильных соединений в мин, x_i	0	1	2	3	4	5	7	
Относительные частоты, m_i/n	8/60	17/60	16/60	10/60	6/60	2/60	1/60	$\Sigma = 60$

Если изучается величина, имеющая непрерывное распределение вероятностей, то возможные значения заполняют целый интервал или всю числовую ось. В этом случае, скорее всего, вариационный ряд не

будет содержать повторяющихся значений. То же самое может иметь место, если наблюдение производится над дискретной случайной величиной, число возможных значений которой очень велико.

Для выборки, в которой нет повторяющихся значений, таблица статистического распределения выборки будет иметь вид (табл. 1.3).

Таблица 1.3

Значения x_i	x_1	x_2	...	x_n
Частоты m_i/n	1/n	1/n	...	1/n

Такая таблица при большом числе наблюдений не содержит полезной информации. В случае, когда вариационный ряд содержит очень много разных значений, прибегают к группировке данных. Обычно группировку стараются провести таким образом, чтобы значения, различие которых для практики незначимо, попали в один и тот же интервал, а те, различия которых уже значимы, попали в разные интервалы.

Группировка состоит в том, что область на оси x, куда попали значения $x_1,...,x_n$, разбивают на интервалы $l_1,...,l_k$ и подсчитывают частоту попадания значений величины в каждый интервал. Самый простой способ группировки – округление данных: сохранить один знак после запятой, округлить до ближайшего целого, до ближайшего, кратного десяти и т.д. Когда эта методика не подходит, прибегают к другим способам. Проще всего взять интервалы одинаковой длины. Число интервалов k следует брать не очень большим, чтобы после группировки ряд не был громоздким, и не очень малым, чтобы не потерять особенности распределения признака. Обычно берут от 6 до 11 интервалов. Согласно формуле Серджеса рекомендуемое число интервалов:

$$k = 1 + 3{,}322 \lg n,$$

Например, так как lg100 = 2, для выборки объема 100 рекомендуемое число интервалов 8. Для выборки объема 50 – 5-6 интервалов.

Величину интервала h можно вычислить по формуле:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n},$$

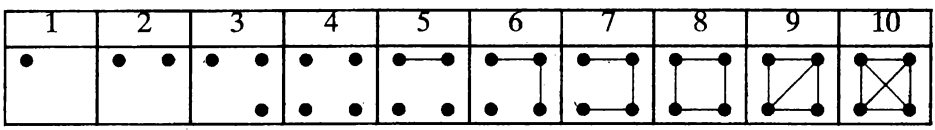
где $x_{max} - x_{min}$ разность между наибольшим и наименьшим значением в выборке (ее размах).

За начало первого интервала рекомендуется брать величину:

$$x_{нач} = x_{min} - 0{,}5h.$$

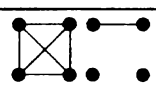
Кроме того, необходимо следить, чтобы не было интервалов, в которые попало меньше 5 значений.

Теперь, просматривая результаты наблюдений, надо определить, сколько значений признака попало в каждый конкретный интервал. При этом в интервал включают значения, большие (или равные) нижней границы интервала и меньшие верхней границы. Частоты попадания в интервалы можно подсчитать следующим образом. Границы интервалов выписывают в столбец. Затем просматривают данные, записанные в том порядке, в котором они были получены. Правее интервала, в который попало данное, ставят точку или черточку. Точки и черточки удобно ставить так, чтобы 10 попаданий образовывали "конверт", для каждого интервала подсчитывается число полных "конвертов" и число точек-черточек в неполном:



Пример 1.5. Число значений, попавших в i-ый интервал, – частоты m_i, которые являются **интервальными частотами,** а отношения m_i/n – **интервальными** или **относительными** (эмпирическими) **интервальными частотами**.

Результаты такой обработки ряда наблюдений можно представить в виде таблицы.

Диаметр ...		Частота m_i	Относительная частота m_i/n
6,67-6,69		2	0,010
6,69-6,71		15	0,075
...	...		
Всего		200	1

Пример 1.6. Получены следующие данные о распределении 100 рабочих цеха по выработке в процентах к предыдущему году:

97,8; 97,0; 101,7; 132,5; ...; 112,3; 104,2; 141,0; 122,1 (100 значений).

$x_{min} = 97$; $x_{max} = 141$.

По формуле Серджеса число интервалов $k = 1 + 3{,}322 \cdot 2 = 8$,

Длина интервала $h = (141 - 97)/8 = 6$,

$x_{нач} = x_{min} - 0{,}5h = 97 - 3 = 94$.

Таблица выработки рабочих в процентах к предыдущему году

i – номер интервала	Выработка в процентах	Количество рабочих (частота)	Доля рабочих (относительная частота)	Накопленная частота	Накопленная относительная частота
1	94-100	3	0,03	3	0,03
2	100-106	7	0,07	3+7=10	0,10
3	106-112	11	0,11	10+11=21	0,21
4	112-118	20	0,20	21+20=41	0,41
5	118-124	28	0,28	41+28=69	0,69
6	124-130	19	0,19	69+19=88	0,88
7	130-136	10	0,10	88+10=98	0,98
8	136-142	2	0,02	98+2=100	1,00

Представленная в таблице **накопленная частота** показывает, сколько наблюдалось вариантов со значением признака, меньшим x. Отношение накопленной частоты к общему числу наблюдений называется **накопленной относительной (эмпирической) частотой.**

Накопленные частоты позволяют с помощью таблицы ответить на вопросы типа: "какова доля рабочих, выработка которых по отношению к прошлогодней меньше 100%?" (ответ: 0,03); или "какова доля тех, у кого выработка увеличилась в 1,3 раза (больше 130%)?" (ответ: 1 – 0,88 = 0,12); "чему равна такая выработка, при которой у половины рабочих выработка меньше, а у половины рабочих больше этого значения?" (ответ: примерно 120%).

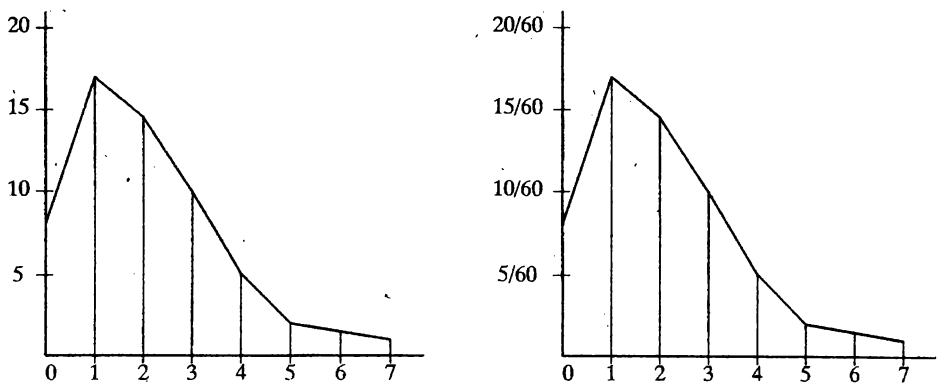
Вариационный ряд, представленный таблицей, построенной с помощью процедуры группировки, будем называть **интервальным** (в отличие от **дискретного ряда**, полученного по выборке из дискретного распределения вероятностей).

1.4. Графическое представление выборки. Полигон, гистограмма, кумулята

Для наглядного представления статистического распределения пользуются графическим изображением вариационных рядов (**полигоном, гистограммой и кумулятой**).

В случае дискретного распределения на оси абсцисс откладывают отдельные значения признака. Из "принимаемых" значений x_i проводят перпендикуляры, длины которых пропорциональны частотам m_i, затем концы соседних перпендикуляров соединяют отрезками прямых. Это **полигон для дискретных вариационных рядов**. Лучше в качестве длин перпендикуляров брать относительные частоты m_i/n. Форма графика сохранится, но мы получаем возможность сравнивать две выборки разного объема.

Пример 1.7. Полигоны числа неправильных соединений в минуту, построенные по таблицам к примеру 1.4.



Гистограмма строится только для интервального вариационного ряда (группированной выборки). На каждом из интервалов значений как на основании, строят прямоугольник с высотой, пропорциональной m_i. Если середины верхних сторон прямоугольников соединить отрезками прямых, а концы этой ломаной еще соединить с серединами соседних интервалов, частоты которых равны 0, а длина равна длине соседнего интервала, то получим **полигон интервального ряда**.

Кумулята – график накопленных частот, сглаженное графическое изображение эмпирической функции распределения.

При построении кумуляты в точке, соответствующей принимаемому значению, для дискретного ряда и в правом конце интервала для

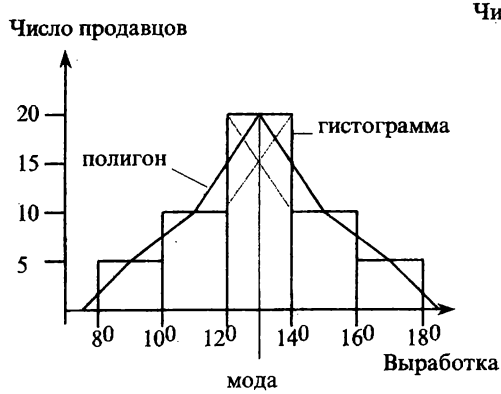
интервального ряда строится перпендикуляр, высота которого пропорциональна накопленной частоте, затем верхние концы перпендикуляров соединяются между собой с помощью прямолинейных отрезков.

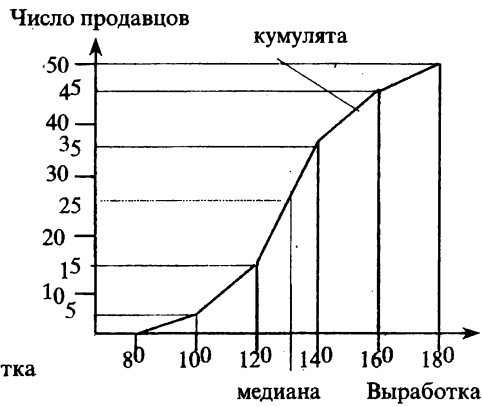
Проще всего показать на конкретном примере, как строятся эти графики.

Пример 1.8.

Таблица распределения продавцов по выработке

Выработка продавцов	Число продавцов	В процентах к итогу	Кумулятивная (накопленная) численность	Накопленная относительная частота
80-100	5	10	5	0,1
100-120	10	20	15(5+10)	0,3
120-140	20	40	35(15+20)	0,7
140-160	10	20	45(35+10)	0,9
160-180	5	10	50(45+5)	1
Итого	50	100		

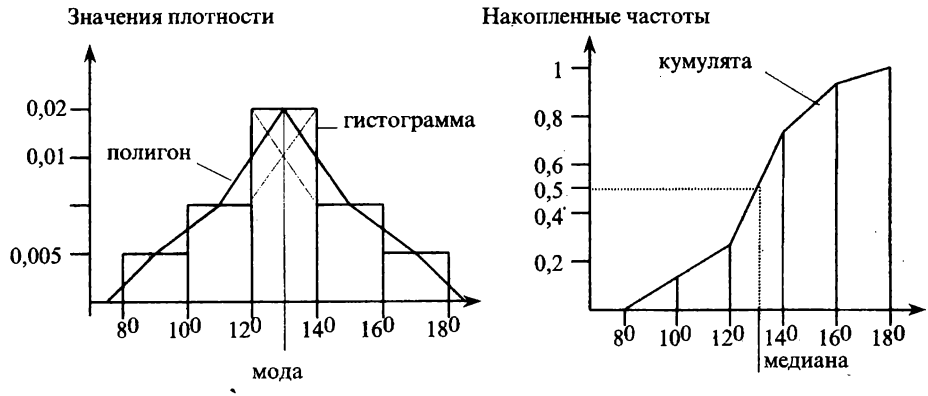




На оси Y могут откладываться не количества, а проценты, или проценты, деленные на константу, например, относительные частоты. Вид графика от этого не изменится.

Так, построенная гистограмма позволяет сравнивать два распределения, имеющие разный объем.

В нашем примере длины интервалов одинаковые. В данном случае при построении гистограммы можно изображать прямоугольники с высотой m_i. Если длины интервалов разные, то при построении гистог-



раммы это надо учитывать. Например, все интервалы имеют длину 10, кроме крайнего, который имеет длину 50 (весь хвост объединен в один интервал). Все попавшие в него данные можно мысленно разбить на 5 одинаковых частей, каждая из которых попала бы в свой интервал длины 10. Следовательно, высота прямоугольника над этим интервалом длины 50 должна браться в 5 раз меньше, чем его частота m, или относительная частота m/n. Лучше всего в качестве высоты прямоугольника брать величину $\frac{m_i}{d_i n}$, где d_i – длина интервала. Часто в таблицах крайние интервалы указываются в форме: "менее x_1" или "свыше x_n". В этом случае их условно заменяют на интервалы той же ширины, что и соседние.

1.5. Числовые характеристики выборки

Рассмотрим некоторые числовые характеристики выборки.

Мода. Для дискретного вариационного ряда легко находится x_m, в котором m имеет наибольшее значение, – это значение, эмпирическая вероятность которого максимальна, называется **модой**. Для интервального ряда легко находится интервал, у которого частота максимальна. Мода находится внутри него. Для вычисления ее значения пользуются формулой линейной интерполяции. На наших гистограммах показано, как она ищется графически. На гистограмме из примера 1.8 выработки продавцов мода равна 130.

Медиана. На графике кумуляты, или сглаженной эмпирической функции распределения, показана эмпирическая медиана. **Медиана** – важная характеристика распределения вероятностей – это середина распределения, т.е. такая точка, что половина принимаемых значений лежит слева от нее, а половина справа. Для дискретного вариационного ряда медиана d ищется по формуле:

$$d = \begin{cases} \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{если n четно} \\ x_{\frac{n+1}{2}}, & \text{если n нечетно} \end{cases}$$

Для **группированной** выборки **медиана** – это точка, в которой площадь гистограммы делится пополам (в примере 1.7 – это такая выработка, когда у 25 продавцов выработка меньше этого числа, а у 25 – больше. Из соображений симметрии видно, что таким значением является число 130). Если медиана лежит практически в центре области принимаемых значений, то это указывает на то, что у выборки нет сильного перекоса вправо или влево, она примерно симметрична относительно медианы. Сдвиг медианы влево (вправо) от центра области принимаемых значений означает больший "вероятностный" удельный вес левой или, соответственно, правой половины принимаемых значений (рис. 1.1).

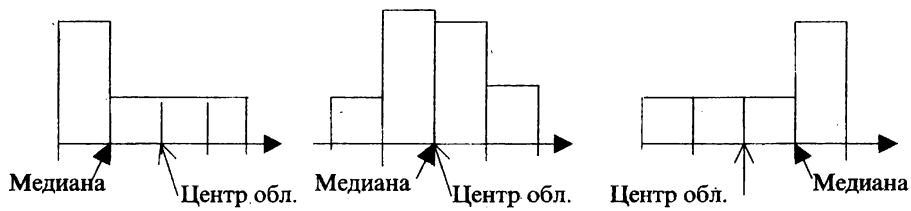


Рис. 1.1

Существуют и другие числовые характеристики выборки. Для их вычисления интервальную таблицу выборки заменяют на дискретную (табл. 1.4). В качестве принимаемых значений указывают середины интервалов группировки.

В дальнейшем будем считать, что и по дискретной и по интервальной выборке задана таблица частот (табл. 1.1) или таблица относительных частот – т.е. таблица эмпирического распределения выборки (табл. 1.2).

В таблице 1.5 приведены формулы, по которым в зависимости от описания данных выборки вычисляются среднее значение и разброс выборки.

Таблица 1.4

Таблица распределения продавцов по выработке (дискретный вариант)

x_i	m_i	m_i / n	Плотность вероятности	Накопленная частота (эмпирическая функция распределения)
90	5	0,1	0,005	0,1
110	10	0,2	0,01	0,3 = (0,1 + 0,2)
130	20	0,4	0,02	0,7 = (0,3 + 0,4)
150	10	0,2	0,01	0,9 = (0,7 + 0,2)
170	5	0,1	0,005	1 = (0,9 + 0,1)
n	50			

Таблица 1.5

	Вариационный ряд задан последовательностью	Задана таблица частот вариационного ряда	Задана таблица относительных частот вариационного ряда
Среднее значение выборки $\bar{x}$	$\frac{1}{n}\sum_{i=1}^{n} x_i$	$\frac{1}{n}\sum_{i=1}^{k} x_j m_j$	$\sum_{j=1}^{k} x_j \frac{m_j}{n}$
Дисперсия (разброс) выборки S^2	$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$	$\frac{1}{n}\sum_{j=1}^{k} x_j^2 m_j - \bar{x}^2 = \frac{1}{n}\sum_{j=1}^{k}(x_j - \bar{x})^2 m_i$	$\sum_{j=1}^{k} x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^{k}(x_j - \bar{x})^2 \frac{m_j}{n}$

Пример 1.9. По выборке 4, 6, 7, 7, 10, 15, 18 (n = 7) найти $\bar{x}$ и S^2.

$\bar{x} = (4 + 6 + 7 + 7 + 10 + 15 + 18)/7 = 9,57.$

$S^2 = 1/7(16 + 36 + 49 + 49 + 100 + 225 + 324) - (9,57)^2 = 114,14 - 91,58 = 22,56$

Пример 1.10. Найти $\bar{x}$ и S^2 по таблице выборки:

Варианты x_i	2	6	12	
Частоты m_i	3	10	7	n = 20

$\bar{x} = (2 \cdot 3 + 6 \cdot 10 + 12 \cdot 7)/20 = 7,5$

$S^2 = 1/20(4 \cdot 3 + 36 \cdot 10 + 144 \cdot 7) - (7,5)^2 = (1/20) \cdot 1380 - 56,25 = 69 - 56,25 = 12,75$

Пример 1.11. Найти $\bar{x}$ и S^2 по таблице выборки:

Варианты x_i	2	6	12	
Частоты m_i/n	0,15	0,5	0,35	1

$\bar{x} = 2 \cdot 0{,}15 + 6 \cdot 0{,}5 + 12 \cdot 0{,}35 = 7{,}5$

$S^2 = (4 \cdot 0{,}15 + 36 \cdot 0{,}5 + 144 \cdot 0{,}35) - (7{,}5)^2 = 69 - 56{,}25 = 12{,}75$

(легко видеть, что примеры 1.2 и 1.3 задают одну и ту же выборку, но в примере 1.2 она задана таблицей частот, а в примере 1.3 – таблицей относительных частот:

0,15 = 3/20; 0,5 = 10/20; 0,35 = 7/20)

1.6. Связь между статистическим распределением выборки и изучаемым распределением вероятностей

Получив выборку и описав ее, мы хотим это описание распространить на всю генеральную совокупность. А генеральная совокупность – это некоторое распределение вероятностей. И наша задача заключается в том, чтобы по полученному экспериментальному материалу сделать выводы о виде распределения или, если есть какие-то теоретические предпосылки о виде распределения, получить оценки значений его числовых параметров – например, если выборка сделана из нормального распределения, оценить с помощью выборки его параметры, или оценить с помощью выборки параметр пуассоновского распределения. Для того, чтобы понять, что нам дает выборка для решения этой задачи, представим себе урну, в которой лежат n шаров. На m_1 из них написано число x_1, на m_2 шарах написано число x_2 и т.д., на m_k шарах – число x_k. Эта урна и есть наша выборка. Согласно классическому определению вероятностей можно говорить о распределении вероятностей этой урны – вероятность достать из нее число x_i равна m_i/n. Заметим, что таблица статистического распределения выборки описывает закон распределения дискретной случайной величины, для которого возможные значения случайной величины – варианты выборки x_i, а соответствующие им вероятности – их относительные частоты m_i/n.

Следовательно, таблица статистического распределения выборки – это таблица распределения выборки, построенного по классической схеме. Эти соображения и послужили основанием для того, чтобы построенную по выборке таблицу значений и их относительных частот назвать таблицей статистического распределения выборки.

Обозначим:

$$\tilde{p}_i = \frac{m_i}{n}$$

В силу того, что $\sum_{i=1}^{k} m_i = n$, выполняется и:

$$\sum_{i=1}^{k} \tilde{p}_i = 1$$

(при округлении этих значений, неизбежном при переводе их в десятичную дробь, следует побеспокоиться, чтобы это правило выполнялось и после округления). В этих обозначениях таблица статистического распределения выборки имеет вид (табл. 1.6):

Таблица 1.6

Эмпирическое распределение (распределение выборки)

Значения x_i	x_1	x_2	...	x_k
Частоты $\tilde{p}_i$	$\tilde{p}_1$	$\tilde{p}_2$	...	$\tilde{p}_k$

1.6.1. Полигон и многоугольник распределения

Если наблюдения производятся над дискретной случайной величиной, то у каждого значения x_i есть теоретическая вероятность его появления в процессе наблюдения. Обозначим ее через p_i. Это означает, что в нашем эксперименте исследуется дискретное распределение, таблица распределения которого имеет вид (табл. 1.7):

Таблица 1.7

Генеральное распределение

Значения x_i	x_1	x_2	...	x_k
Частоты p_i	p_1	p_2	...	p_k

Таким образом, мы хотим вместо распределения, задаваемого табл. 1.7, на практике пользоваться распределением, задаваемым табл. 1.6. Так что встает вопрос об оценке близости этих распределений.

Значение $\tilde{p}_i$ является выборочным аналогом (он вычисляется по выборке) вероятности p_i появления значения x_i. В силу теоремы Бернулли, вычисляемая по выборке относительная частота $\tilde{p}_i$ обладает свойством статистической устойчивости и стремится по вероятности к вероятности p_i. Поясним коротко, что означает термин "сходится по вероятности". В курсе анализа изучается понятие сходимости. Последовательность $\{a_n\}$ называется сходящейся к a при n, стремящемся к бесконечности, если за счет роста n можно добиться, чтобы разность $|a_n - a|$ стала как угодно мала. **Сходимость случайной величины по вероятности к некоторому значению** означает, что, несмотря на увеличение числа испытаний n, могут встретиться значения случайной величины, довольно сильно отличающиеся от предельного значения, но процент таких испытаний будет с ростом n уменьшаться (вероятность отклонения от предела стремится к 0). Сходимость $\tilde{p}_i$ по вероятности к p_i означает, что для любого $\varepsilon > 0$, несмотря на рост n, будут встречаться выборки, для которых нарушается соотношение $|\tilde{p}_i - p_i| < \varepsilon$, но с ростом числа испытаний n вероятность (доля или процент таких выборок среди множества всех возможных выборок) стремится к нулю.

Полигон, построенный по относительным частотам, – это просто статистический (эмпирический) многоугольник генерального распределения. В силу теоремы Бернулли при n, стремящемся к бесконечности, он сходится по вероятности к многоугольнику генерального распределения.

1.6.2. Гистограмма и плотность вероятности

В случае группировки данных, если n увеличивать и длины интервалов группировки уменьшать, то гистограмма и полигон неограниченно приближаются (в каждой точке **сходятся по вероятности**) к кривой плотности вероятности случайной величины.

Действительно, при построении гистограммы будем строить прямоугольники с высотой $m_i/d_i n$, где d_i – длина интервала I_i. Вспомним, что для **непрерывно распределенной случайной величины** ξ вероятность попасть в интервал (a,b) вычисляется по формуле:

$$P\{a < \xi < b\} = \int_a^b f(x)dx,$$

где f(x) – плотность распределения вероятностей величины ξ. Следовательно, вероятность попасть в интервал $[x,x+\Delta x)$ выражается через плотность f(x), если длина интервала Δx мала, следующим образом:

$$P\{x \le \xi < x + \Delta x\} \approx f(x)\Delta x.$$

Из этой формулы видно, что плотность вероятности – это вероятность, "приходящаяся в данной точке на единицу измерения". Если эмпирическая вероятность попасть в i-ый интервал равна m_i/n, а d_i – это длина i-ого интервала, то эмпирическая вероятность, приходящаяся на единицу измерения, равна m_i/d_in. Если строить прямоугольники с такими высотами, то суммарная площадь всех прямоугольников равна 1. Так, построенная гистограмма изображает эмпирическую плотность. При n, стремящемся к бесконечности, она в каждой точке сходится по вероятности к теоретической плотности.

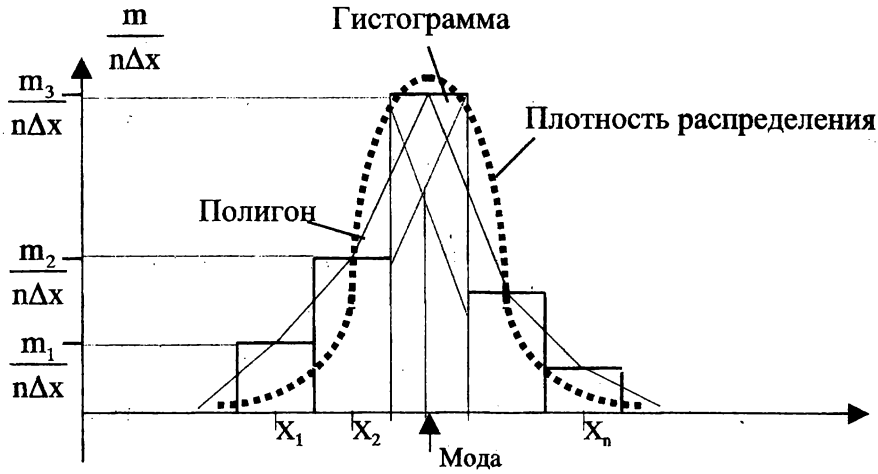


Рис. 1.2

По виду построенной нами гистограммы (рис. 1.2) можно предположить, что она построена по выборке из нормального распределения.

Приведенная ниже гистограмма дает основание полагать, что выборка получена из равномерного распределения (рис. 1.3).

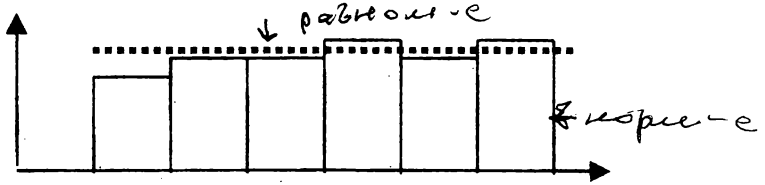


Рис. 1.3

На рис. 1.4 приведена еще одна гистограмма – не из нормального и не из равномерного распределения. По ее виду можно предположить, что выборка сделана из показательного распределения.

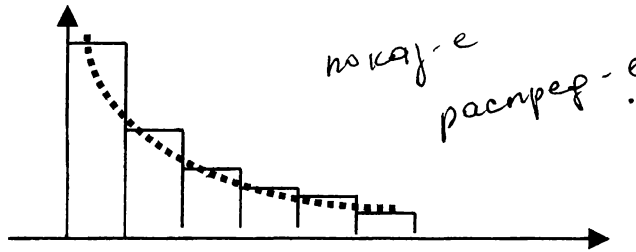


Рис. 1.4

Эти примеры демонстрируют, как по гистограмме, построенной по выборке, можно примерно оценить тип распределения вероятностей. В дальнейшем мы научимся более точно решать задачу проверки по выборке гипотезы о генеральном распределении.

1.6.3. Кумулята, эмпирическая и теоретическая функции распределения

По выборочным данным можно вычислить эмпирическую вероятность события ($\xi < x$) для любого x (она равна $\nu(x)/n$, где $\nu(x)$ – число вариант, меньших x) – т.е. найти эмпирическую функцию распределения. Введенные нами ранее "накопленные частоты" – это и есть значе-

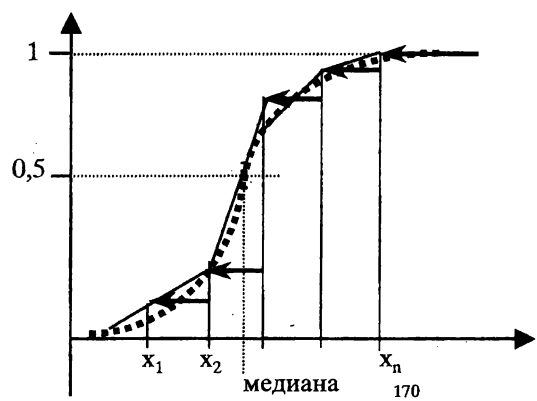


Рис. 1.5

ния эмпирической функции распределения в концах интервалов группировки, **кумулята** – ее сглаженное графическое изображение.

Можно показать, что кумулята в каждой точке ***сходится по вероятности*** к теоретической функции распределения.

Остается добавить, что вследствие той же теоремы Бернулли числовые характеристики эмпирического распределения (математическое ожидание, дисперсия, мода, медиана и многие другие) в широком классе случаев также сходятся по вероятности к соответствующим характеристикам генерального распределения.

Таким образом, теорема Бернулли дает теоретическое обоснование для основной рекомендации, которую предлагает математическая статистика, – надо составить описание выборки, вычислить ее параметры и приписать свойства выборки всей совокупности, а параметрами, вычисленными по выборке, оценить параметры всей генеральной совокупности. При большом объеме выборки ошибка будет небольшой. В силу того, что изучается объект, в природу которого заложен элемент случайности, все оценки, вычисляемые по выборке, являются случайными значениями. Две разные выборки, полученные из одной и той же генеральной совокупности, дадут разные статистические таблицы, будут иметь разные средние значения, разные дисперсии и т.д. И хотя возможны выборки, которые на самом деле дают существенное отличие от теоретического распределения, с ростом объема выборки процент таких "плохих" выборок стремится к нулю. Кроме теоретического обоснования, что такой подход возможен, необходимо получить представление о точности и надежности эмпирической оценки. Это и будет предметом нашего рассмотрения.

Таким образом, первая задача, которую мы будем решать, – это задача построения оценок для параметров генеральной совокупности по данным выборки и изучение их точности и надежности. То есть мы считаем, что известен вид теоретического (генерального) распределения, но не известны и подлежат оценке параметры этого распределения. Так, например, для биномиального распределения по результатам n раз проведенного эксперимента надо оценить значение p. Или известно, что интересующая нас величина распределена нормально, над ней n раз проводятся испытания; по результатам испытаний надо построить оценки для ее математического ожидания и среднеквадратического отклонения (или дисперсии). Вспомним, что нормальное распределение целиком определяется двумя параметрами – математическим ожиданием и дисперсией. Нормальное распределение является одним из самых распространенных распределений вероятности. Следовательно, данная задача охватывает очень большой круг приложений.

Эта же методика позволит решить и задачу сравнения 2-х выборок. Например, сделать вывод о совпадении или различии средних значений их распределений. Эту задачу надо решить, например, в таких ситуациях, как: "внесены исправления в статью закона, повлияло ли это на среднее количество преступлений по этой статье?", или "изменился режим работы, повлияло ли это на производительность труда?" и т. д.

2. ПОСТРОЕНИЕ ТОЧЕЧНЫХ ОЦЕНОК ДЛЯ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.1. Распределения вероятностей, зависящие от параметра

Итак, предположим, что заранее известен вид теоретического распределения интересующего нас признака ξ $F(x,\Theta)$, где Θ – параметр распределения. Это означает, что в непрерывном случае выборка производится из распределения, для которого задана плотность $f(x,\Theta)$ (непрерывная модель), а в дискретном случае – вероятности $P\{\xi=x_i\} = f(x_i,\Theta)$, когда количество принимаемых значений x_i конечно или счетно (дискретная модель). Значение параметра Θ (или, если распределение зависит от нескольких параметров, значения параметров) этого распределения не известно, и должна быть найдена по данным выборки оценка для него (или для них).

При работе с выборочными данными мы будем строить функции, зависящие от выборочных значений $x_1,\ldots,x_n$. Любую функцию $\Theta_n(x_1,\ldots,x_n)$, зависящую от выборки и поэтому являющуюся случайной величиной, принято называть **статистикой**.

Если в качестве оценки параметра предлагается число – точка на координатной оси, то оценка называется **точечной**.

2.2. Метод моментов

Очень часто параметры распределения вероятностей являются моментами распределения (или функциями от них). Моменты являют-

Таблица 2.1

	Дискретное распределение	Непрерывное распределение
Начальный момент порядка l	$a_l = \sum x_i^l p_i$	$a_l = \int x^l f(x)dx$
Центральный момент порядка l	$b_l = \sum (x_i - a_1)^l p_i$	$b_l = \int (x - a_1)^l f(x)dx$

ся важными вероятностными характеристиками распределения. Напомним формулы, по которым они вычисляются (табл. 2.1).

"Центр", относительно которого вычисляется центральный момент – это первый начальный момент a_1, суммирование выполняется по всем принимаемым значениям, интегрирование – по всей области определения.

Первый начальный момент a_1 – это математическое ожидание распределения, второй центральный момент b_2 – это дисперсия.

В силу важности математического ожидания и дисперсии напомним их определение и важнейшие свойства, которыми мы будем в дальнейшем пользоваться в наших рассуждениях (табл. 2.2, 2.3).

Таблица 2.2

Определение математического ожидания и дисперсии

	Дискретное распределение	Непрерывное распределение
Математическое ожидание $M(\xi) = a$	$\sum_{i=1}^{n} x_i \cdot p_i$	$\int_{-\infty}^{\infty} x \cdot f(x)dx$
Дисперсия $D(\xi) = \sigma^2$	$\sum_{i=1}^{n} (x_i - a)^2 \cdot p_i = \sum_{i=1}^{n} x_i^2 \cdot p_i - a^2$	$\int_{-\infty}^{\infty} (x - a)^2 \cdot f(x)dx = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - a^2$

Чтобы описать разброс, рассеяние случайной величины, применяют разные характеристики, чаще всего, дисперсию. Но, поскольку размерность **дисперсии случайной величины** ξ равна квадрату размерности самой случайной величины, применяют даже не дисперсию $D(\xi)$, а **среднеквадратическое (стандартное) отклонение** – корень из дисперсии $\sigma = \sqrt{D(\xi)}$. Эта характеристика имеет ту же размерность, что и сама случайная величина.

Из свойств дисперсии немедленно следует:

$$\sigma(C) = 0\ ;\ \sigma(C \cdot \xi) = |C| \cdot \sigma(\xi);\ \sigma(\xi \pm \eta) = \sqrt{\sigma^2(\xi) + \sigma^2(\eta)}\ .$$

Таблица 2.3

Свойства математического ожидания и дисперсии

	Математическое ожидание	Дисперсия
Случайная величина – константа	$M(C) = C$	$D(C) = 0$
Случайная величина умножается на константу	$M(C\cdot\xi) = C\cdot M(\xi)$	$D(C\cdot\xi) = C^2\cdot D(\xi)$
Случайные величины складываются	$M\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} M(\xi_i)$, в частности $M(\xi+C)=M(\xi)+C$	$D\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} D(\xi_i)$ (для независимых слагаемых), в частности $D(\xi-\eta) = D(\xi)+D(\eta)$ и $D(\xi+C) = D(\xi)$

Моменты – очень важные характеристики распределения. Они много проще функции распределения – это просто числа, но знание их дает очень много информации о распределении. Математическое ожидание – первый начальный момент, это средняя точка распределения, тот "центр", вокруг которого все распределено и вычисляются центральные моменты. Дисперсия – второй центральный момент, характеризует разброс вокруг математического ожидания (среднего). Третий центральный момент, деленный на третью степень среднеквадратического отклонения, – асимметрия распределения. Если он не равен нулю, то распределение несимметрично относительно своего центра. Четвертый центральный момент, деленный на четвертую степень среднеквадратического отклонения, – эксцесс, характеризует "плосковершинность" распределения и т. д.

Мы сосредоточим свое внимание на математическом ожидании и дисперсии распределения, как самых важных его характеристиках.

В качестве примера распределений, зависящих от параметра, приведем самые распространенные законы распределения и их числовые характеристики. Связь параметров распределения с его числовыми характеристиками станет очевидной.

Пример 2.1. Испытание с двумя исходами, биномиальное распределение. Пусть в результате испытания с вероятностью p происходит событие A (случайная величина ζ – индикатор наступления A

приняла значение 1), а с вероятностью q = 1–p противоположное ему событие $\bar{A}$ (ζ приняла значение 0). Это распределение можно задать таблицей:

x_i	0	1
p_i	q	p

$$M(\zeta) = 0 \cdot q + 1 \cdot p = p$$

$$D(\zeta) = 0 \cdot q + 1^2 \cdot p - p^2 = p \cdot (1-p) = pq$$

Такое распределение полностью определяется параметром p – математическим ожиданием. С этим распределением тесно связано биномиальное распределение – распределение числа успехов, полученных при n независимых испытаниях, проводящихся над такой случайной величиной (схема Бернулли). Вероятность того, что случайная величина, имеющая биномиальное распределение (число успехов в схеме Бернулли), примет значение m, задается **формулой Бернулли**:

$$p_n(m) = C_n^m \cdot p^m \cdot q^{n-m} = \frac{n \cdot (n-1) \cdot \ldots \cdot (n-m+1)}{1 \cdot 2 \cdot \ldots \cdot m} p^m q^{n-m}$$

$$M(\xi) = M\left(\sum_{k=1}^{n} \zeta_k\right) = \sum_{k=1}^{n} M(\zeta_k) = np,$$

$$D(\xi) = D\left(\sum \zeta_k\right) = \sum_{k=1}^{n} D(\zeta_k) = npq.$$

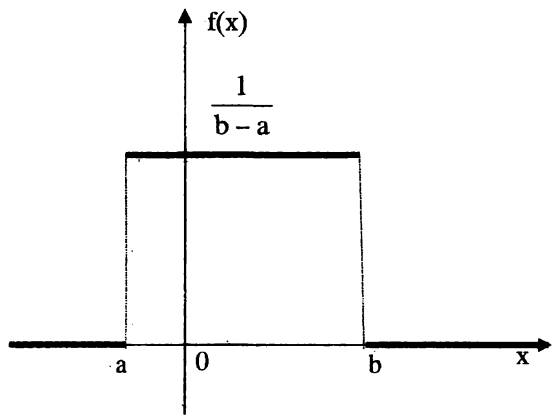
Пример 2.2. Распределение Пуассона. Случайная величина, которая принимает значение m с вероятностью $P_m(\lambda) = \frac{\lambda^m e^{-\lambda}}{m!}$, где m = 0,1,2,..., а λ – положительная постоянная величина, называется **распределенной по закону Пуассона с параметром** λ. Из курса теории вероятностей известно, что

$$M(\xi) = D(\xi) = \lambda.$$

Примерами величины, распределенной по закону Пуассона, являются число новорожденных в сутки, число аварий и т.д. Очень важным свойством закона Пуассона и его параметра λ является "воспроизво-

димость": сумма двух независимых случайных величин, распределенных по Пуассону с параметрами λ_1 и λ_2, распределена также по закону Пуассона с параметром $\lambda_1+\lambda_2$; параметр λ случайных событий, протекающих во времени и распределенных по закону Пуассона, пропорционален времени, он равен λt (λ – это среднее число событий, наступающих в некоторую единицу времени). Следовательно, если мы знаем значение параметра λ для промежутка времени t_1, мы, тем самым, знаем его и для любого другого промежутка t_2 (это будет $\lambda t_2/t_1$). Для вычисления пуассоновских вероятностей разработаны таблицы.

Пример 2.3. Равномерное распределение. Непрерывная случайная величина является равномерно распределённой на интервале [a,b], если её плотность вероятности равна константе на этом интервале и нулю вне его. График плотности вероятностей, ее выражение, математическое ожидание и дисперсия приведены ниже:



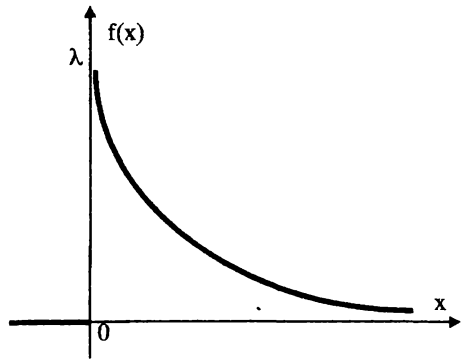
$$f(x)=\begin{cases}\dfrac{1}{b-a}, x\in[a,b]\\ 0 \quad , x\notin[a,b]\end{cases};$$

$$M(\xi)=\frac{1}{b-a}\int_a^b x\,dx=\frac{b+a}{2}$$

$$D(\xi)=\frac{1}{b-a}\int_a^b (x-M)^2\,dx=\frac{(b-a)^2}{12}$$

Равномерное распределение дает пример распределения, зависящего от двух параметров; его параметры a и b связаны с моментами распределения функциональной зависимостью.

Пример 2.4. Показательное распределение. Непрерывная случайная величина имеет показательное распределение, если её плотность вероятности имеет вид, как на приведенном графике:



$$f(x)=\begin{cases}\lambda\cdot e^{-\lambda\cdot x}, x\in[0,\infty); & M(\xi)=\lambda\int_0^{\infty}xe^{-\lambda x}dx=\frac{1}{\lambda};\\ 0 \quad , x\in(-\infty,0); & D(\xi)=\lambda\int_0^{\infty}(x-\frac{1}{\lambda})^2e^{-\lambda x}dx=\frac{1}{\lambda^2}.\end{cases}$$

Интегралы берутся по частям. Таким образом, показательное распределение зависит от параметра λ, а его математическое ожидание равно 1/λ.

Пример 2.5. Нормальное распределение. Нормальное распределение – распределение Гаусса – играет особую роль в теории вероятностей и её приложениях. Это наиболее часто встречающийся на практике закон распределения. Этому закону подчиняется, при соблюдении определённых условий, распределение суммы достаточно большого числа случайных величин, каждая из которых может иметь произвольное распределение. Нормальное распределение задается плотностью f(x):

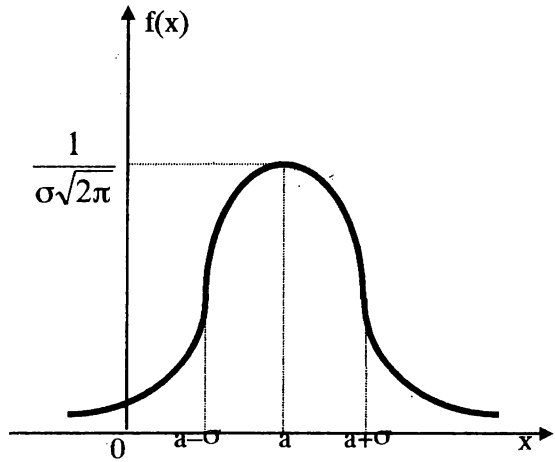
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-a)^2}{2\sigma^2}}$$

Определение математического ожидания и дисперсии приводит к вычислению соответствующих интегралов, что дает:

$$M(\xi)=a,\ D(\xi)=\sigma^2,\ \sigma(\xi)=\sigma.$$

Таким образом, параметрами, определяющими нормальное распределение, являются a – математическое ожидание и σ – среднеквадратическое отклонение (корень из дисперсии), для его обозначения употребляется запись N(a, σ).

График плотности нормального распределения



Из приведенных примеров видно, как важно научиться строить оценки именно для моментов.

Для построения точечных оценок для математического ожидания и дисперсии генеральной совокупности или любых других моментов вспомним, что таблица статистического распределения выборки задает распределение выборки. В качестве оценок характеристик генерального распределения мы решили брать значения тех же характеристик, но вычисленных для распределения выборки. Для распределения выборки, которое является обычным дискретным распределением, заданным таблицей, можно сосчитать любые моменты (так как они вычисляются по выборке, мы их будем называть эмпирическими). Их и возьмем в качестве оценок теоретических моментов. Если оцениваемый параметр является функцией от моментов распределения, то в эту функцию вместо неизвестных теоретических значений моментов подставим эмпирические значения.

Выбирая в качестве оценок такие статистики, мы воспользовались методикой, предлагаемой *методом моментов*. Этот метод впервые был использован К.Пирсоном в 1894 г.

Метод моментов – метод получения оценок параметров, который состоит в том, что если оцениваемый параметр распределения является функцией от моментов распределения (в самом простом случае сам является моментом), то в эту функцию просто подставляются эмпирические значения моментов, а полученное значение берется в качестве оценки для параметра.

Самый простой пример применения метода моментов. Математическое ожидание – первый начальный, а дисперсия – второй центральный момент. В качестве оценок для их генеральных значений мы возьмем первый начальный и второй центральный моменты выборки (эмпирические моменты). Они обладают свойствами математического ожидания и дисперсии, описанными в табл. 2.3.

2.3. Вычисление эмпирических моментов

Выборке соответствует дискретное распределение (табл. 2.4).

Таблица 2.4

Значения x_i	x_1	x_2	...	x_k
Частоты $\tilde{p}_i$	$\tilde{p}_1$	$\tilde{p}_2$	...	$\tilde{p}_k$

В случае, если выборка задана вариационным рядом, эта таблица имеет следующий вид (табл. 2.5):

Таблица 2.5

Значения x_i	x_1	x_2	...	x_n
Частоты $\tilde{p}_i$	1/n	1/n	...	1/n

Вычисление моментов эмпирического распределения согласно табл. 2.6 производится по следующим формулам (во всех формулах n – объем выборки):

Таблица 2.6

	Вариационный ряд общего вида	Вариационный ряд задан таблицей
Начальный эмпирический момент порядка l	$a_l = \frac{1}{n}\sum_{i=1}^{n} x_i^l$	$a_l = \sum_{j=1}^{k} x_j^l \frac{m_j}{n} = \sum_{j=1}^{k} x_j^l \tilde{p}_j$
Центральный эмпирический момент порядка l	$b_l = \frac{1}{n}\sum_{j=1}^{n} (x_j - \bar{x})^l$	$b_l = \sum_{j=1}^{k} (x_j - \bar{x})^l \frac{m_j}{n} = \sum_{j=1}^{k} (x_j - \bar{x})^l \tilde{p}_j$

Сравнение с табл. 1.5 показывает, что первый начальный момент или математическое ожидание выборки a_1 – это введенное нами ранее среднее значение выборки, которое мы обозначили через $\overline{x}$. Выборочную дисперсию b_2 мы назвали разбросом выборки и обозначили S^2.

Таким образом, согласно методу моментов оценкой для математического ожидания генерального распределения надо взять $\overline{x}$, оценкой для дисперсии – S^2. Формулы для их расчета содержатся в табл. 1.5.

Пример 2.6. Число неправильных соединений на телефонной станции имеет пуассоновское распределение. В течение часа каждую минуту фиксировали число неправильных соединений, имевших место в эту минуту. Полученные данные представлены в таблице (см. пример 1.4). Оценить среднее число неправильных соединений в час.

Решение. На языке теории вероятностей по результатам выборки надо получить оценку математического ожидания, то есть параметра λ распределения Пуассона. Согласно методу моментов в качестве оценки для среднего числа неправильных соединений в минуту надо взять:

$$\overline{x} = (0 \cdot 8 + 1 \cdot 17 + 2 \cdot 16 + 3 \cdot 10 + 4 \cdot 6 + 5 \cdot 2 + 7 \cdot 1)/60 = 120/60 = 2$$

Согласно правилу устойчивости распределения Пуассона, упомянутому нами ранее, среднее число неправильных соединений в час в 60 раз больше среднего числа неправильных соединений в минуту. Следовательно, оценкой для среднего числа неправильных соединений в час является значение 2·60 = 120.

2.4. Свойства точечных оценок

Для того чтобы оценка неизвестного параметра, т.е. статистика $\Theta(x_1,\ldots,x_n)$ (коротко мы ее будем обозначать Θ_n), давала хорошее приближение неизвестного параметра Θ распределения генеральной совокупности $F(x,\Theta)$, желательно, чтобы она удовлетворяла следующим требованиям.

Математическое ожидание оценки параметра по всевозможным выборкам данного объёма должно равняться истинному значению определяемого параметра:

$$M(\Theta_n) = \Theta$$

В этом случае **оценку** называют **несмещённой**. В противном случае оценку называют смещённой.

Если это требование не выполняется, то оценка Θ_n, полученная по разным выборкам, будет в среднем либо завышать, либо занижать зна-

чение Θ. К сожалению, часто практичести важные оценки являются смещенными, хотя и слабо. Для оценки, смещенной слабо:

$$M(\Theta_n) \to \Theta \text{ (при n, стремящемся к } \infty).$$

Было бы ошибочным считать, что несмещенная оценка всегда дает хорошее приближение оцениваемого параметра. Чем больше дисперсия оценки, тем больше ее значения рассеяны вокруг ее среднего значения и, следовательно, удалены от значения оцениваемого параметра. По этой причине к оценке предъявляется требование эффективности.

Оценка называется **эффективной**, если при заданном объеме выборки n она имеет наименьшую дисперсию.

При увеличении объёма выборки оценка должна сходиться по вероятности к истинному значению параметра; в этом случае **оценку** называют **состоятельной**.

Для состоятельных оценок значительные ошибки при оценивании маловероятны.

Если дисперсия несмещенной оценки при n, стремящемся к бесконечности, стремится к нулю, то такая оценка оказывается и состоятельной. Это непосредственно вытекает из неравенства Чебышева:

$$P(|\Theta_n - \Theta| < \varepsilon) \geq 1 - \frac{\sigma^2_{\Theta_n}}{\varepsilon^2}.$$

Из неравенства Чебышева видно, что, для того чтобы доказать несмещенность и состоятельность оценки, достаточно изучить ее математическое ожидание и дисперсию. Следует заметить, что на практике часто для простоты расчетов используют незначительно смещенные оценки или оценки, обладающие большей дисперсией по сравнению с эффективными оценками.

Оценки метода моментов обычно состоятельны, но не являются "наилучшими" в смысле эффективности. Тем не менее метод моментов часто используется на практике, так как требует сравнительно простых вычислений.

2.5. Понятие надежности оценки

Метод моментов дает нам точечные (числовые) оценки для вероятности p в схеме испытания с двумя исходами (значение m/n), для параметра λ пуассоновского распределения, для среднего a (выборочное среднее $\overline{x}$), для дисперсии σ^2 (выборочная дисперсия S^2).

Обсудим вопрос об их точности и надежности.

Статистическая оценка Θ_n является лишь приближенным значени-

ем неизвестного параметра Θ даже в том случае, если она несмещенная (в среднем совпадает с Θ), состоятельная (стремится к Θ с ростом n) и эффективная (обладает наименьшей степенью случайных отклонений от Θ).

В силу случайной природы изучаемых характеристик их сходимость к предельным значениям – сходимость по вероятности. Это означает, что точность оценок, вычисленных по выборке, имеет место не "всегда", а только для подавляющего числа выборок. Таким образом, кроме обычного понятия точности оценок встает вопрос еще и об их надежности – в каком проценте случаев точность оценки не нарушается.

В силу этого с точностью оценки, полученной на основе выборки, математическая статистика связывает понятие "уровня доверия" к ней. "Уровень доверия"– это и есть ее надежность, процент (доля) случаев, для которых гарантируется требуемая точность оценки. То есть точности оценки можно доверять не на все 100%, а лишь с некоторым "уровнем доверия". Например, если указано, что уровень доверия для оценки 0,95, то из 100 выборок примерно 5 дадут оценки, которые на самом деле не удовлетворяют требованиям точности. Является конкретная выборка "плохой" или "хорошей", к сожалению, сказать нельзя, так что если делать на основе выборочного метода вывод о всей совокупности, то вероятность ошибиться остается. Математическая статистика дает методику вычисления этой вероятности.

Описание точности и надежности оценки Θ_n параметра Θ дает распределение вероятностей разности оценки и истинного значения – $(\Theta_n - \Theta)$. Оно задает среднее значение ошибки и позволяет оценить вероятность слишком большого отклонения оценки от истинного значения. Покажем на примере выборочного среднего $\overline{x}$, какие выводы его распределение вероятностей позволяет сделать о точности и надежности точечной оценки $\overline{x}$ для генерального среднего a.

2.6. Распределение выборочного среднего

Оценка генерального среднего – выборочное среднее $\overline{x}$ – является случайной величиной, значение которой зависит от того, какие значения приняли варианты x_i. Если наблюдения проводятся над нормальной случайной величиной с параметрами a и σ, то как сумма нормально распределенных случайных величин она подчиняется нормальному закону. Найдем ее математическое ожидание и дисперсию. Воспользуемся для этого приведенными в табл. 2.3 свойствами математического ожидания и дисперсии:

$$M\bar{x} = M\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n}\sum_{i=1}^{n} Mx_i = \frac{an}{n} = a$$

$$D\bar{x} = D\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \frac{1}{n^2} D\sum_{i=1}^{n} x_i = \frac{1}{n^2}\sum_{i=1}^{n} Dx_i = \frac{\sigma^2}{n}$$

Следовательно, у величины $\bar{x}$ то же математическое ожидание a, что и у генерального распределения, а дисперсия в n раз меньше: $\sigma_{\bar{x}}^2 = \sigma^2/n$.

Эти соотношения выведены без учета требования нормальности генерального распределения. В силу центральной предельной теоремы, если число наблюдений n велико, то каким бы ни было распределение у случайной величины, из которой делается выборка, если у него существует дисперсия, выборочное среднее $\bar{x}$, являясь суммой большого числа случайных величин, подчиняется закону, близкому к нормальному, так что формула

$$\bar{X}_n \approx N\left(a, \frac{\sigma}{\sqrt{n}}\right)$$

верна в достаточно широком классе случаев. Попутно мы получили доказательство того, что выборочное среднее является *несмещенной* и, в силу теоремы Чебышева, *состоятельной* оценкой.

Замечание. Оценка S^2 смещена. Исправить эмпирическую дисперсию S^2, чтобы получить несмещенную точечную оценку s^2 для неизвестной дисперсии надо следующим образом:

$s^2 = \frac{n}{n-1} S^2$ – несмещенная оценка для дисперсии.

2.7. Связь между точностью и надежностью оценки

Для того, чтобы оценить точность и надежность точечной оценки для генерального среднего, воспользуемся тем, что статистика $\bar{x}$ имеет нормальное распределение с математическим ожиданием a и среднеквадратическим отклонением $\frac{\sigma}{\sqrt{n}}$.

Напомним о некоторых свойствах величины $\xi \sim N(a,\sigma)$.

Выполним над величиной ξ операцию, которая называется **нормировкой.** Вычтем из нее a и поделим эту разность на σ. Из свойств математического ожидания и дисперсии следует, что математическое ожидание новой случайной величины будет равно 0, а ее среднеквадратическое отклонение – единице:

$$M(\xi)=a;\quad D(\xi)=\sigma^2;\quad M\left(\frac{\xi-a}{\sigma}\right)=0;\quad D\left(\frac{\xi-a}{\sigma}\right)=1.$$

Величину $\xi\sim N(0,1)$ с нулевым средним и единичной дисперсией называют стандартной нормальной. Ее плотность вероятности задается формулой:

$$f_\xi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Как для $\xi\sim N(0,1)$, так и для $\xi\sim N(a,\sigma)$, вероятность попадания в любой интервал может быть выражена через функцию Лапласа $\Phi(x)$:

$$\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{-x}^{x}e^{-\frac{t^2}{2}}dt=\frac{2}{\sqrt{2\pi}}\int_{0}^{x}e^{-\frac{t^2}{2}}dt.$$

Для нее разработаны таблицы. В нашем руководстве это табл. 4 файла материалов. В силу очевидного из анализа соотношения $\Phi(-x)=-\Phi(x)$ значения функции Лапласа в таблице заданы только для положительных x.

$$P\left(A\le\xi\le B;\xi\square N(0,1)\right)=$$

$$=\frac{1}{\sqrt{2\pi}}\int_{A}^{B}e^{-\frac{t^2}{2}}dt=\frac{1}{\sqrt{2\pi}}\int_{A}^{0}e^{-\frac{t^2}{2}}dt+\frac{1}{\sqrt{2\pi}}\int_{0}^{B}e^{-\frac{t^2}{2}}dt=\frac{1}{2}\left(\Phi(B)-\Phi(A)\right),$$

$$P\left(A\le\xi\le B;\xi\square N(a,\sigma)\right)=$$

$$=\frac{1}{\sigma\sqrt{2\pi}}\int_{A}^{B}e^{-\frac{(t-a)^2}{2\sigma^2}}dt=\frac{1}{\sqrt{2\pi}}\int_{\frac{A-a}{\sigma}}^{\frac{B-a}{\sigma}}e^{-\frac{y^2}{2}}dy=\frac{1}{2}\left[\Phi\left(\frac{B-a}{\sigma}\right)-\Phi\left(\frac{A-a}{\sigma}\right)\right]$$

(сделали замену переменных $\frac{t-a}{\sigma}=y;\quad dy=\frac{dt}{\sigma}$).

В частности для интервала, симметричного относительно математического ожидания a:

$$p(|\xi - a| \le k\sigma) = p(a - k\sigma \le \xi \le a + k\sigma) =$$

$$= \frac{1}{2}\left[\Phi\left(\frac{a + k\sigma - a}{\sigma}\right) - \Phi\left(\frac{a - k\sigma - a}{\sigma}\right)\right] =$$

$$= \frac{1}{2}[\Phi(k) - \Phi(-k)] = \frac{1}{2} \cdot 2 \cdot \Phi(k) = \Phi(k)$$

Из этой формулы вытекает так называемое **правило трех** σ (рис. 2.1).

$$p(|\xi - a| \le 3\sigma) = p(a - 3\sigma \le \xi \le a + 3\sigma) = \Phi(3) = 0{,}9973.$$

То есть, *практически достоверно то, что нормально распределённая величина примет значение, отличающееся от её математического ожидания по модулю не более чем на 3σ, иначе говоря, "практически невозможно" появление значения, выходящего за пределы этого интервала.*

Последнее обстоятельство находит широкое применение в различных приложениях.

Понятие "практически невозможно" и противоположное ему понятие "практически достоверно" часто (особенно в экономике) использу-

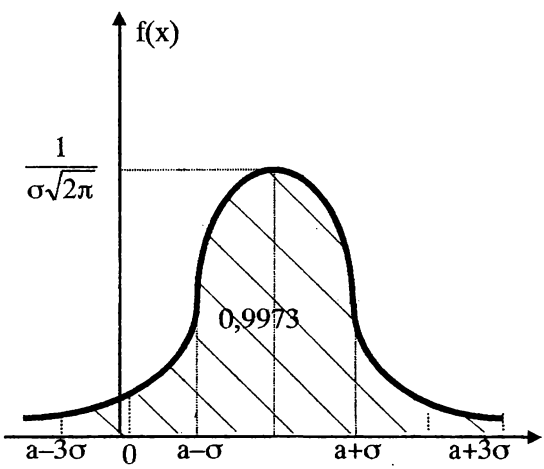


Рис. 2.1

ются в смягченном варианте, когда "практически достоверными" считаются уже события, вероятность которых не 0,997, а 0,95. То есть в экономике обычно используется "правило двух σ", так как

$$p(|\xi - a| \le 2\sigma) = \Phi(2) = 0{,}9544.$$

Введем обозначение k_β для корня уравнения:

$$P\{|\xi| < k_\beta\} = P\{-k_\beta < \xi < k_\beta\} = \Phi(k_\beta) = \beta, \text{ где } \xi = N(0,1).$$

Для случайной величины $\xi \sim N(a,\sigma)$ так же, как мы выводили правило трех σ, можно вывести формулу:

$$p\{|\xi - a| < k_\beta\sigma\} = \Phi(k_\beta) = \beta$$

(в частности, для $\beta = 0{,}95$ $k_{0,954} \approx 2$ – она обращается в правило двух σ, а для $\beta = 0{,}997$ $k_{0,997} = 3$ – в правидо трех σ).

Пример 2.7. Затаривание мешков с сахаром производится без систематических ошибок. Случайные ошибки подчинены нормальному закону со среднеквадратическим отклонением σ = 200 г и, вследствие отсутствия систематической ошибки, математическим ожиданием 0. Найти вероятность того, что затаривание будет проведено с ошибкой, не превосходящей по абсолютной величине 400 г.

Решение. В задаче рассматривается случайная величина – ошибка взвешивания, то есть разность (ξ – a) между случайным значением веса мешка сахара и его нормативным значением a – математическим ожиданием. Требуется найти:

$$p(|\xi - a| < 400) = p(a - 400 < \xi < a + 400) =$$

$$= \frac{1}{2}\left[\Phi\left(\frac{a + 400 - a}{\sigma}\right) - \Phi\left(\frac{a - 400 - a}{\sigma}\right)\right] = \Phi\left(\frac{400}{200}\right) = \Phi(2) = 0{,}954.$$

Это типичная задача курса теории вероятностей.

Легко видеть, что одновременно мы решили и следующую задачу: какую максимальную ошибку затаривания мы можем гарантировать с вероятностью 0,95? *Ответ:* 2σ = 400 г (лишь у 5% мешков ошибка веса окажется больше). А если мы хотим определить максимальную ошибку недовеса-перевеса, в которую укладываются не 95%, а "почти все" мешки (99.7% мешков), то какова она? *Ответ:* 3σ = 600 г.

Заодно мы решили и задачу о том, как обеспечить требуемую точность: какую среднеквадратическую ошибку должен иметь автомат, проводящий затаривание, чтобы с вероятностью 0,95 ошибка перевеса-недовеса не превосходила 100 г? *Ответ:* 50 г.

Вторая и третья постановка вопроса характерна для задач математической статистики. В задачах математической статистики требуется не по интервалу находить вероятность попадания в него, а *по заданной вероятности ищется интервал, обладающий некоторыми свойствами, в который с заданной вероятностью* β *попадает случайная величина.* Для $\xi \sim N(0,1)$, если ищется симметричный относительно 0 интервал, в который с вероятностью β попадает ξ, этот интервал равен $[-k_\beta, k_\beta]$).

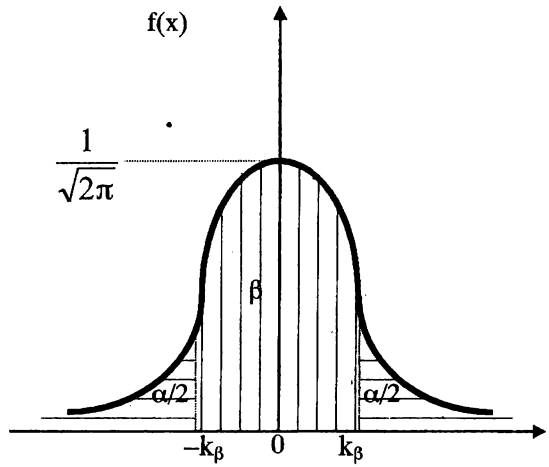


Рис. 2.2

Тем самым, для определения надежности оценки $\bar{x}$ мы пользуемся свойствами и таблицами нормального распределения. Если мы хотим иметь надежность, равную β, то в таблице нормального распределения находим число k_β, такое что:

$$P(|\xi| < k_\beta) = P(-k_\beta < \xi < k_\beta) = \Phi(k_\beta) = \beta$$

(при этом надо помнить, что в данном случае по значению функции ищется значение аргумента).

Например, $k_{0,95} = 1{,}96$; $k_{0,997} = 3$ (вспомните правило 2σ и 3σ).

Если требуемая надежность равна 0,9; 0,95; 0,98; 0,99 или 0,997 (а это самые распространенные на практике значения), то проще воспользоваться табл. 5 файла материалов, задающей значения k_β нормального распределения для перечисленных значений β.

Итак, в силу того, что $\bar{x} \sim N\left(a, \frac{\sigma}{\sqrt{n}}\right)$, и $\frac{\bar{x}-a}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, имеем:

$$P\left\{\left|\frac{\bar{x}-a}{\frac{\sigma}{\sqrt{n}}}\right| < k_\beta\right\} = P\left\{|\bar{x}-a| < k_\beta \frac{\sigma}{\sqrt{n}}\right\} = \beta$$

Эта формула описывает точность, с которой значение $\bar{x}$ описывает генеральное среднее a и надежность этой оценки (уровень доверия к ней).

3. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ДЛЯ ПАРАМЕТРОВ

3.1. Понятие доверительного интервала

Эта же схема применяется во всех рассуждениях, когда строятся интервальные оценки для параметров. Для описания интервальной оценки разработана специальная терминология. А именно, говорят: для параметра Θ построен **доверительный интервал надежности (или уровня доверия)** β. При построении этого интервала мы исходим из соображения, что если в процессе эксперимента для статистики получено некоторое значение, то оно принадлежит области (будем называть ее I_β), вероятность которой близка к 1 (равна β). Эту вероятность называют **доверительной вероятностью.**

Обычно в качестве области I_β берут интервал, накрывающий значение оцениваемого параметра с вероятностью β. Его и называют доверительным интервалом с уровнем доверия β. **Доверительный интервал строится в соответствии с распределением вероятностей используемой статистики.**

Величина уровня доверия β, как мы видели, влияет на величину интервала: чем больше уровень доверия, тем шире интервал.

Мы можем решить для себя, на какой риск мы готовы пойти в нашем конкретном случае, а математическая статистика даст точность оценки, гарантируемую для заданного допустимого риска. Или, наоборот, получив от математической статистики ответ, что для данной точности уровень доверия меньше допустимого, мы можем постараться добиться результатов, заслуживающих большего доверия. Например, увеличить число объектов, участвующих в исследовании. Математическая статистика показывает, что чем больше число отобранных для обследования объектов, тем меньше вероятность ошибки, и дает функциональную зависимость между объемом выборки, точностью и вероятностью ошибки. При этом предлагаются "оптимальные" методики, при использовании которых величина вероятности ошибки минимальна.

Выбрав уровень риска, на который мы готовы пойти, мы в каком-то смысле вместо достоверных событий, вероятность которых равна 1, начинаем считать за "практически достоверные" события, вероятность которых только близка к 1 (степень близости к 1 и есть наш уровень риска). Таким образом, можно сказать, что математическая статистика предлагает методики, следуя которым, мы будем не ошибаться в своих рассуждениях не "всегда", а только "практически всегда", в соответствии с выбранным нами "уровнем доверия" (указанием, что мы подразумеваем под понятием "практически всегда"). Принято уровень доверия брать равным 0,95 или 0,99. Если, приняв уровень доверия 0,99, мы будем по выборкам строить доверительные интервалы, то в среднем 1 из 100 интервалов не будет содержать истинное значение параметра (какой именно 1 из 100 мы, конечно, не можем сказать). Если примем уровень доверия 0,95 и будем по выборкам строить доверительные интервалы, то в среднем 5 из 100 интервалов не будут содержать истинное значение параметра. Выбор уровня доверия остается за нами. Если цена ошибки высока (разорение, смертельный исход при операции) – может быть, следует задать уровень доверия равным 0,999, если ошибка грозит тем, что придется взять кредит в банке – можно удовольствоваться уровнем 0,95. Если лекарство безвредно, то достаточно того, что "оно помогает с уровнем доверия 70%", чтобы рекомендовать его для применения. Доверительные интервальные оценки вычисляются в соответствии с выбранным уровнем доверия. При этом, конечно, надо учитывать, что чем выше заказанный уровень доверия, тем более расплывчатым будет ответ. Ответы математическая статистика выдает в виде формул, в которые уровень доверия входит как параметр. Так что часто они позволяют выбрать стратегию, позволяющую добиться желательной точности с нужным уровнем доверия к результатам.

3.2. Доверительный интервал для среднего в случае, когда среднеквадратическое отклонение σ теоретического распределения известно

Легко видеть, что выведенная нами формула:

$$\bar{x} - k_\beta \frac{\sigma}{\sqrt{n}} < a < \bar{x} + k_\beta \frac{\sigma}{\sqrt{n}}$$

– это формула **доверительного интервала с уровнем доверия** β **для математического ожидания a нормального распределения для случая, когда известно среднеквадратическое отклонение распределения** σ. При взгляде на нее ясно, что чем больше n, тем

уже интервал, а чем более близкую гарантию β мы требуем, тем доверительный интервал шире. Кроме того, она позволяет оценить, каков должен быть объем выборки n, чтобы точность оценки, полученной по ней для генерального среднего, не превосходила заданного значения ε (эпсилон) с уровнем доверия β. В случаях, когда определение объема выборки в нашей власти, мы можем вычислить, на сколько надо увеличить ее объем, чтобы добиться нужной точности. Так как точность обратно пропорциональна корню из n, то для того, чтобы повысить точность в 2 раза, объем выборки надо увеличить в 4 раза; чтобы повысить точность в 10 раз, число испытаний надо увелить в 100 раз.

Написанное соотношение было выведено в предположении, что дисперсия исходного распределения известна (например, она раньше была установлена по выборке объема больше 50). А целью данного эксперимента является оценить только среднее. Рассмотрим теперь более распространенный случай – когда не только среднее, но и дисперсия генеральной совокупности не известны.

3.3. Доверительный интервал для среднего в случае, когда среднеквадратическое отклонение σ теоретического распределения неизвестно

Нами было показано, что когда дисперсия известна, выборочное среднее имеет нормальное распределение с параметрами a и $\frac{\sigma}{\sqrt{n}}$. Мы отнормировали его – вычли из него его математическое ожидание и поделили полученную разность на его среднеквадратическое отклонение. Тем самым перешли от него к стандартной нормальной величине и воспользовались ее свойствами и таблицами.

Заменим в этой операции неизвестное среднеквадратическое отклонение его эмпирической оценкой. Рассмотрим статистику $\tau = \frac{\sqrt{n}(\bar{x} - a)}{s} = \frac{\sqrt{n-1}(\bar{x} - a)}{S}$ (в этой формуле s – корень из s^2 – исправленной, несмещенной оценки для дисперсии: $s^2 = \frac{n}{n-1} S^2$).

При нахождении распределения вероятностей для статистики τ мы должны учесть, что неизвестное среднеквадратическое отклонение мы заменили в формулах на его эмпирический аналог. Можно показать, что τ имеет распределение Стьюдента с (n–1) степенями свободы.

Распределением Стьюдента с n степенями свободы называется распределение случайной величины

$$t = \frac{\xi}{\sqrt{\frac{1}{n}\cdot\chi_n^2}} = \frac{\xi}{\sqrt{\frac{1}{n}\cdot\sum_{i=1}^{n}\xi_i^2}},$$

где ξ, ξ_1, ... , ξ_n – независимые, стандартные нормальные случайные величины. Это распределение симметрично, а при значениях $n > 20$ практически неотличимо от нормального. При меньших n разница все-таки есть и ее надо учитывать. В задачах экономики в методе скользящей средней (при прогнозировании по данным за четыре квартала ожидаемого значения для пятого квартала и во многих других) используются выборки небольшого объема (так называемые малые выборки). В этих задачах используется распределение Стьюдента с небольшим числом степеней свободы. Для распределения Стьюдента существуют многочисленные статистические таблицы.

Применив те же рассуждения, которые мы применяли при построении доверительного интервала для среднего при известной дисперсии, получаем формулу для доверительного интервала для среднего в случае неизвестной дисперсии. А именно, обозначим через $t_{n;\beta}$ значение, для которого

$$P\left\{-t_{n;\beta} < \xi < +t_{n;\beta}\right\} = \beta,$$

где ξ имеет распределение Стьюдента с n степенями свободы. Значение $t_{n;\beta}$ по заданному значению β находится по таблицам распределения Стьюдента (табл. 6 файла материалов) аналогично тому, как мы искали k_β для нормального распределения, используя табл. 5. При этом надо, правда, еще учесть значение n, что не должно вызвать затруднений. Как правило, таблицы распределения Стьюдента задаются не для всех β, а только для наиболее употребительных значений 0,9, 0,95 и 0,99. *Если n велико (больше 20), и под рукой нет таблиц распределения Стьюдента, а имеются только более распространенные таблицы нормального распределения, то можно воспользоваться ими, считая, что с хорошей точностью* $t_\beta = k_\beta$. Например, если требуемый уровень доверия 0,95, то можно взять $t_{n;\beta} = 2$, а если уровень доверия 0,997, то $t_{n;\beta} = 3$ (правило 2σ и 3σ для нормального распределения).

Таким образом, для статистики τ, имеющей распределение Стьюдента с (n–1) степенью свободы, можно записать:

$$P\{|\tau| < t_{n-1;\beta}\} = \beta$$

и, проделав простые тождественные преобразования, получаем, что с вероятностью β выполняется

$$\bar{x} - t_{n-1;\beta}\frac{s}{\sqrt{n}} < a < \bar{x} + t_{n-1;\beta}\frac{s}{\sqrt{n}} \text{ или } \bar{x} - t_{n-1;\beta}\frac{S}{\sqrt{n-1}} < a < \bar{x} + t_{n-1;\beta}\frac{S}{\sqrt{n-1}}$$

Это формула **доверительного интервала с уровнем доверия β для математического ожидания а нормального распределения для случая, когда среднеквадратическое отклонение распределения σ неизвестно**.

Пример 3.1. Для проверки фасовочной установки были отобраны и взвешены 20 упаковок. Были получены следующие результаты (в граммах): 246; 247; 247,3; 247,4; 251,7; 252,5; 252,6; 252,8; 252,8; 252,9; 253; 253,6; 254,6; 254,7; 254,8; 256,1; 256,3; 256,8; 257,4; 259,2.

Найти доверительный интервал для математического ожидания с надёжностью 0,95, предполагая, что измеряемая величина распределена нормально.

Решение. Находим точечные оценки а и σ:

$$\tilde{a} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{20}\sum_{i=1}^{20} x_i = 252{,}98$$

$$\tilde{\sigma}^2 = s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{1}{19}\sum_{i=1}^{20}(x_i - \bar{x})^2 = 13{,}3$$

$$\tilde{\sigma} = s = 3{,}65$$

Определяем по таблице распределения Стьюдента (табл. 6 файла материалов) для доверительной вероятности $\beta = 0{,}95$ и числу степеней свободы $(n - 1) = 19$ соответствующее значение $t_\beta = 2{,}093$ и по формуле находим искомый интервал:

$$252{,}98 - \frac{2{,}093 \cdot 3{,}65}{\sqrt{20}} \le a \le 252{,}98 + \frac{2{,}093 \cdot 3{,}65}{\sqrt{20}} \text{ или } 251{,}27 \le a \le 254{,}69.$$

3.4. Оценка требуемого объема выборки

Формулы доверительного интервала позволяют заодно решить еще одну интересную задачу: каков должен быть объем выборки n, чтобы с надежностью β точность оценки, полученной по ней для а, не превосходила заданного значения ε, то есть $|\bar{x} - a| < \varepsilon$ (среднеквадратическое отклонение известно)?

Действительно, так как по формуле доверительного интервала с

вероятностью β выполняется $|\bar{x} - a| < k_\beta \frac{\sigma}{\sqrt{n}}$, то нужное n находится из уравнения $k_\beta \frac{\sigma}{\sqrt{n}} = \varepsilon$, то есть $n = \left(\frac{k_\beta \sigma}{\varepsilon}\right)^2$. Следовательно, результат тем точнее, чем больше объем выборки.

3.5. Доверительный интервал для вероятности успеха в схеме Бернулли

Пусть проводятся независимые испытания, в которых событие A наступает с неизвестной вероятностью p. Ставится задача с помощью выборочных испытаний построить для p точечную и интервальную оценки. Метод моментов нам указывает, что в качестве точечной оценки надо взять значение $\tilde{p} = \frac{m}{n}$, где m – число успехов. Действительно, таблица статистического распределения выборки объема n, в которой m раз произошел "успех" (выпала 1) и n–m раз "неуспех" (выпал 0), имеет вид (табл. 3.1):

Таблица 3.1

X_i	0	1
$\tilde{p}_i$	(n-m)/n	m/n

Неизвестная вероятность p равна математическому ожиданию генерального распределения. Следовательно, метод моментов рекомендует нам взять в качестве оценки для p эмпирическое среднее:

$$\bar{x} = 1 \cdot \frac{m}{n} + 0 \cdot \frac{n-m}{n} = \frac{m}{n}.$$

Воспользуемся тем, что по теореме Муавра-Лапласа величина $\frac{m - np}{\sqrt{npq}}$ распределена приблизительно нормально, т.е.:

$$P\left\{\left|\frac{m - np}{\sqrt{npq}}\right| < k_\beta\right\} \cong \Phi(k_\beta) = \beta.$$

Отсюда

$$P\left\{\left|m - np\right| < k_\beta \cdot \sqrt{npq}\right\} = P\left\{\left|\frac{m}{n} - p\right| < k_\beta \sqrt{\frac{pq}{n}}\right\} = \beta.$$

Воспользовавшись тем, что q = 1–p, получаем, что с вероятностью β выполняется неравенство:

$$\tilde{p} - k_\beta \sqrt{\frac{p(1-p)}{n}} \le p \le \tilde{p} + k_\beta \sqrt{\frac{p(1-p)}{n}}.$$

Таким образом, для построения доверительного интервала для p можно воспользоваться таблицами нормального распределения.

Границы интервала зависят от неизвестной величины p. В руководствах по статистике можно найти формулы для границ, лишенные этого недостатка; мы же воспользуемся тем, что при больших n неизвестное p можно заменить его эмпирическим значением:

$$\tilde{p} - k_\beta \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + k_\beta \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}.$$

Формула доверительного интервала позволяет заодно решить еще одну задачу: каков должен быть объем выборки n, чтобы с надежностью β точность оценки, полученной по ней для p, не превосходила заданного значения ε, т.е. $\left|\tilde{p} - p\right| < \varepsilon$?

Действительно, по формуле доверительного интервала с вероятностью β выполняется неравенство $\left|\tilde{p} - p\right| < k_\beta \frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{n}}$, т.е. результат тем точнее, чем больше объем выборки. Нужное n находится из уравнения $k_\beta \frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{n}} = \varepsilon$, т.е. $n = \frac{k_\beta^2}{\varepsilon^2}\tilde{p}(1-\tilde{p})$.

Замечание. Для бесповторной выборки (выборки без возвращения) из генеральной совокупности объема N длина доверительных интервалов меньше:

$$\tilde{p} - k_\beta \sqrt{\frac{m/n(1-m/n)}{n}} \sqrt{1-\frac{n}{N}} \le p \le \tilde{p} + k_\beta \sqrt{\frac{m/n(1-m/n)}{n}} \sqrt{1-\frac{n}{N}}.$$

Такую же поправку следует сделать для случая бесповторной выборки и в формулах доверительных интервалов для среднего значения нормального распределения (уменьшить дисперсию в (1–n/N) раз.

3.6. Односторонние доверительные интервалы

На практике часто пользуются односторонними доверительными интервалами. Например, страховой компании не страшно, если произойдет страховых случаев намного меньше среднего, но страшно, если их произойдет намного больше среднего. Оценивая при покупке среднюю доходность объекта, лучше оценить ее по формуле "не меньше, чем"; при изучении среднего уровня воды в реке в областях, подверженным наводнениям, интересуются уровнем, выше которого вода не поднимется, а в областях, подверженных засухе, наоборот – уровнем, ниже которого вода не опустится.

В этом случае строят не симметричный относительно оценки интервал, а максимально расширяют его за счет одной из его границ. Если мы построим двусторонний доверительный интервал с гарантией β, а затем максимально расширим его в одну сторону, то получим односторонний интервал с большей гарантией $\beta' = \beta+(1-\beta)/2=(1+\beta)/2$ (рис. 3.1). Например, если $\beta = 0{,}90$, то $\beta' = 0{,}90 + 0{,}10/2 = 0{,}95$, а если $\beta = 0{,}95$, то $\beta' = 0{,}95 + 0{,}05/2 = 0{,}975$. Таким образом, "односторонний" подход позволяет увеличить уровень доверия, вернее, вдвое снизить ошибку $\alpha = 1-\beta$ (или при том же уровне доверия сузить интервал – вместо t_β можно взять $t_{2\beta-1}$). Если при построении двусторонних доверительных интервалов надо было решать уравнение:

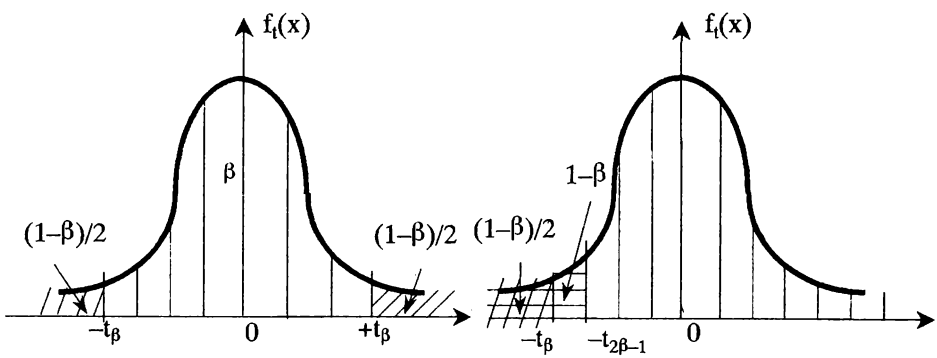


Рис. 3.1

$$\int_{-t_\beta}^{+t_\beta} f(x)dx = \beta,$$

то для построения односторонних доверительных интервалов надо решать уравнение:

$$\int_{-\infty}^{-t_\alpha} f(x)dx = \alpha = 1-\beta,$$

$$\int_{t_\alpha}^{\infty} f(x)dx = \alpha = 1-\beta.$$

Плотность f(x), как для нормального распределения, так и для распределения Стьюдента, симметричная функция (f(x) = f(–x)), следовательно, для них ошибка, состоящая в непопадании в интервал, симметричный относительно математического ожидания, делится поровну между попаданием в полуинтервал $[-\infty,-t_\beta]$ и полуинтервал $[t_\beta,\infty]$, то есть вероятность каждого такого полуинтервала вдвое меньше ошибки двустороннего интервала. Следовательно, для поиска нужного k_β при построении одностороннего интервала, надо по заданному уровню доверия β найти его ошибку $\alpha = 1-\beta$, затем удвоить ее, взять $\alpha' = 2\alpha$, вычислить для нее новый уровень доверия $\beta' = 1-\alpha'$ и найти $k_{\beta'}$ для двустороннего интервала с таким уровнем доверия. Таблицы 5 и 6 файла материалов используют это обстоятельство и содержат значения половинных ошибок, позволяющих прямо по ним получать нужное значение (см. ниже). В табл. 6 (см. файл материалов) одно и то же t соответствует ошибке двустороннего интервала, задаваемой в верхней строке таблицы и ошибке одностороннего интервала, задаваемой в нижней строке таблицы (она вдвое меньше). В табл. 5, менее перегруженной информацией из-за отсутствия параметра n (число степеней свободы), для значения ошибки одностороннего интервала отведен свой столбец (стоит отметить, что ошибка одностороннего интервала вдвое меньше ошибки двустороннего интервала).

Очень часто статистические таблицы составляются именно для односторонних интервалов. Этот способ является универсальным, а для несимметричных распределений единственно возможным. Значения u_p, для которых выполняется $\int_{-\infty}^{u_p} f(x)dx = P$, называются **квантилями**.

Обозначим через F(x) – функцию распределения стандартного нор-

мального закона. Она связана с функцией Лапласа Φ(x) простым соотношением:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{0} + \int_{0}^{x} = 0.5 + 0.5 \cdot \Phi(x) = 0.5[1 + \Phi(x)].$$

Как для ξ~N(0,1), так и для η~N(a,σ), вероятность попадания в любой полуинтервал может быть выражена через F(x), то есть вычислена с помощью таблиц функции F(x). Действительно:

$$P(\xi \le B) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{B} e^{-\frac{t^2}{2}} dt = F(B)$$

$$P(\eta \le B) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{B} e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{B-a}{\sigma}} e^{-\frac{y^2}{2}} dy = F\left(\frac{B-a}{\sigma}\right)$$

(мы сделали замену переменных $\frac{t-a}{\sigma} = y; \quad dy = \frac{dt}{\sigma}$).

$$P(\xi > B) = \frac{1}{\sqrt{2\pi}} \int_{B}^{\infty} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{\infty} - \int_{-\infty}^{B} = 1 - F(B)$$

$$P(\eta > B) = \frac{1}{\sigma\sqrt{2\pi}} \int_{B}^{\infty} e^{-\frac{(t-a)^2}{2\sigma^2}} dt = \frac{1}{\sqrt{2\pi}} \int_{\frac{B-a}{\sigma}}^{\infty} e^{-\frac{y^2}{2}} dy = 1 - F\left(\frac{B-a}{\sigma}\right).$$

Отсюда для ξ~N(a,σ):

$$P\{\xi < a - u_\beta\sigma\} = P\{\xi > a + u_\beta\sigma\} = \frac{1}{2}[1 - \Phi(u_\beta)] = 1 - F(u_\beta) = 1 - \beta = \alpha,$$

где через u_β обозначена квантиль нормального распределения.

В частности:

$$P\{\xi < a - 3\sigma\} = P\{\xi > a + 3\sigma\} = 1/2[1 - 0{,}9973] = 0{,}00135,$$

$$P\{\xi < a - 2\sigma\} = P\{\xi > a + 2\sigma\} = 0{,}023,$$

$$P\{\xi > a - 2\sigma\} = P\{\xi < a + 2\sigma\} = 0{,}977.$$

Приведем формулу, которая нам понадобится для вычисления одностороннего доверительного интервала с уровнем доверия 0,95:

$$P\{\xi < a-1{,}65\sigma\} = P\{\xi > a+1{,}65\sigma\} = 0{,}05,$$

$$P\{\xi > a-1{,}65\sigma\} = P\{\xi < a+1{,}65\sigma\} = 0{,}95.$$

Ниже мы рассмотрим примеры и на односторонние интервалы. А пока приведем формулы для односторонних доверительных интервалов, аналогичные формулам двусторонних интервалов:

1. Односторонние доверительные интервалы для математического ожидания a нормального распределения с уровнем доверия β для случая, когда среднеквадратическое отклонение распределения σ известно:

$$-\infty < a < \overline{x} + k_\beta \frac{\sigma}{\sqrt{n}}$$

и

$$\overline{x} - k_\beta \frac{\sigma}{\sqrt{n}} < a < \infty$$

(k_β отыскивается в табл. 5 файла материалов по $\alpha = 1-\beta$, находя его значение в 3-м столбце).

2. Односторонние доверительные интервалы для математического ожидания μ нормального распределения с уровнем доверия β для случая, когда среднеквадратическое отклонение распределения σ неизвестно:

$$-\infty < a < \overline{x} + t_{n-1,\beta} \frac{s}{\sqrt{n}}$$

и

$$\overline{x} - t_{n-1,\beta} \frac{s}{\sqrt{n}} < a < \infty$$

($t_{n-1;\beta}$ отыскивается в таблице 6 файла материалов по $\alpha = 1-\beta$, находя его значение в нижней строке).

Аналогично выписываются формулы односторонних доверительных интервалов для вероятности p схемы Бернулли.

4. ПРОВЕРКА СТАТИСТИЧЕСКОЙ ГИПОТЕЗЫ

4.1. Проверка гипотезы о равенстве среднего числовому значению

Понятие "статистическая гипотеза" более емкое, чем просто оценка значения неизвестного параметра. Пусть с помощью статистического эксперимента мы хотим проверить гипотезу о том, что неизвестное среднее μ равно некоторому значению μ_0. Эта гипотеза будет основной. При этом альтернативной ей может быть гипотеза $\mu \neq \mu_0$ или более сложная гипотеза типа $\mu < \mu_0$ или $\mu > \mu_0$. Например, известно, что в среднем за смену на станке производится 110 деталей. Станок сломался и его "отремонтировали". Получив на "отремонтированном" станке показатели за n смен, мы хотим проверить гипотезу: "производительность станка не изменилась", когда альтернативной гипотезой является, что она изменилась, или что производительность станка увеличилась, или что производительность станка уменьшилась.

При решении таких задач также применяется аппарат построения для соответствующей статистики области I_β, вероятность попадания в которую β достаточно близка к 1. При попадании статистики, построенной по выборке, в эту область принимается основная гипотеза (в нашем примере, что производительность станка не изменилась и равна 110); в противном случае, если значение статистики попало в область, противоположную I_β, принимается альтернативная гипотеза (производительность станка изменилась). В задачах о проверке гипотез принято **область**, противоположную I_β, называть **критической**, а число $\alpha = 1-\beta$ – **уровнем значимости**. Уровень значимости α обычно берут равным 0,05, иногда 0,01. При $\alpha = 0{,}05$ мы, проверяя на деле истинную гипотезу о том, что $\mu = \mu_0$, будем ее отбрасывать с вероятностью 0,05, т.е. в среднем 5 из 100 истинных гипотез. В простых случаях областями I_β оказываются уже знакомые нам доверительные интервалы. При проверке гипотезы $\Theta = \Theta_0$ мы строим с доверительной вероятностью $\beta = 1-\alpha$ при альтернативной гипотезе $\Theta \neq \Theta_0$ двусторонний, а при гипотезах $\Theta > \Theta_0$ и $\Theta < \Theta_0$ односторонние с нижней границей $x_н$ и верхней границей $x_в$ доверительные интервалы. Если этот интервал покрывает Θ_0, гипотеза H_0 принимается, если не накрывает – отвергается. Приведем некоторые примеры.

Пример 4.1. В задаче про "ремонт" станка проверяем гипотезу об изменении производительности станка, если за 31 смену получены данные о производительности станка, для которых $\overline{x} = 100$, $s^2 = 20^2 = 100$. Уровень значимости $\alpha = 0{,}05$ ($\beta = 0{,}95$) и $\nu = n - 1 = 30$. Значения $t_{n;\beta}$,

участвующие в построении доверительного интервала, отыскиваются в таблице 6, β или α в верхней строке:

$$I_{0,95} = \bar{x} \pm t_{30;0,95} \frac{20}{\sqrt{30}} = 100 \pm 2,04 \cdot 3,65 = 100 \pm 7,45$$

Вывод. Гипотеза о том, что производительность станка не изменилась, не проходит на уровне значимости 5%, так как старая производительность, равная 110, в 95-процентный доверительный интервал, построенный по новой средней производительности, не попала. Более того, она не попала бы в доверительный интервал, даже если бы мы задались 98-процентным уровнем доверия, для которого $t_{30;0,98} = 2,46$ ($I_{0,98} = 100 \pm 8,98$). Т.е. наша выборка показала, что гипотеза о том, что производительность станка не изменилась, не проходит даже на уровне значимости 2%. Только при уровне доверия 0,99 ($t_{30;0,98} = 2,75$) интервал становится таким большим ($I_{0,99} = 100 \pm 10,04$), что мы уже не можем быть на 99% уверены, что изменение выработки не случайно. Увидев, что новые показатели хуже старых, берем в качестве альтернативной гипотезу о том, что новое среднее меньше старого (такая альтернатива естественна, если $\bar{x} < \mu_0$) то есть, что производительность станка уменьшилась. Это предположение подтверждается даже на уровне значимости 0,01. Действительно, строим односторонний доверительный интервал для уровня доверия 0,99. Значения $t_{n;\beta}$, участвующие в построении одностороннего доверительного интервала, отыскиваются в таблице 6, β или α в нижней строке:

$$I_{0,99} = (-\infty, \bar{x} + t_{v;0,01} \frac{20}{\sqrt{30}}) = (-\infty, 100 + 2,46 \cdot 3,65) = (-\infty, 108,98).$$

Так как $\mu_0 = 110$ не входит в построенный односторонний интервал, можно принять гипотезу о том, что производительность уменьшилась, на уровне значимости 1%.

Перечислим критерии, по которым, не привлекая понятия доверительного интервала, проверяется статистическая гипотеза о том, что среднее значение генеральной совокупности $\mu = \mu_0$ на уровне значимости α (они выведены из формул для двустороннего и одностороннего доверительного интервала для уровня доверия $\beta = 1-\alpha$).

Вычисляем по выборке значение статистики $T = \dfrac{\bar{x} - \mu_0}{\dfrac{s}{\sqrt{n}}}$.

1. Критическая область для односторонней проверки гипотезы, что среднее значение генеральной совокупности $\mu = \mu_0$ по сравнению с альтернативой $\mu \neq \mu_0$, на уровне значимости α определяется неравенством:

$$|T| > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в верхней строке).

2. Критическая область для односторонней проверки гипотезы, что среднее значение генеральной совокупности $\mu = \mu_0$ по сравнению с альтернативой $\mu > \mu_0$, на уровне значимости α определяется неравенством:

$$T > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке)

3. Критическая область для односторонней проверки гипотезы, что среднее значение генеральной совокупности $\mu = \mu_0$ по сравнению с альтернативой $\mu<\mu_0$, на уровне значимости α определяется неравенством:

$$T < -t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

Если вычисленное значение статистики T попадает в критическую область, то основная гипотеза отвергается. Вероятность попадания в эту область равна принятому уровню значимости α. В этом случае принимается альтернативная гипотеза.

В нашем примере про станок $T = -2{,}74$, а $t_{30;0,01} = 2{,}58$, так что основная гипотеза не проходит, а проходит альтернативная гипотеза $\mu < 110$ при уровне значимости 0,01.

Использование доверительного интервала для параметра в задачах проверки гипотезы о его значении имеет то преимущество, что для случая, когда основная гипотеза не проходит, этот метод сразу дает эмпирическую оценку параметра. В качестве значения параметра конкурирующей гипотезы часто берут эмпирическое значение параметра (в частности, в качестве a можно взять $\bar{x}$). В нашем случае такой оценкой для новой производительности станка будет число 100.

4.2. Сравнение двух генеральных средних

Рассмотрим две независимые выборки $x_1,x_2,\ldots,x_n$ и $y_1,y_2,\ldots,y_n$, извлеченные из нормальных генеральных совокупностей с одинаковыми дисперсиями $\sigma_x^2 = \sigma_y^2 = \sigma^2$, причем объемы выборок соответственно n и m, а средние μ_x,μ_y и дисперсия σ^2 неизвестны. Требуется проверить гипотезу о том, что $\mu_x = \mu_y$. Альтернативной является гипотеза $\mu_x \neq \mu_y$.

Как известно, выборочные средние – нормально (или приблизительно нормально) распределенные величины, следовательно, их разность $\bar{x} - \bar{y}$ – нормальная величина со средним $\mu_x - \mu_y$ и дисперсией, которая вычисляется по формуле:

$$D(\bar{x} - \bar{y}) = D\bar{x} + D\bar{y} = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{\sigma^2(m+n)}{mn}.$$

Если бы дисперсия σ^2 была известна, мы могли бы для проверки гипотезы воспользоваться свойствами и таблицами нормального распределения, как мы это делали при построении доверительного интервала для среднего при известной дисперсии. В силу того, что σ^2 неизвестна, заменим в наших рассуждениях неизвестную дисперсию на ее эмпирический аналог.

Итак, для проверки гипотезы $\mu_x = \mu_y$ построим статистику:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где

$$s^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{m}(y_i - \bar{y})^2\right] = \frac{1}{n+m-2}(nS_x^2 + mS_y^2).$$

Теперь к статистике t применим те же рассуждения, которые мы применяли к статистике T.

Если гипотеза $\mu_x = \mu_y$ верна, статистика t имеет распределение Стьюдента с n+m–2 степенями свободы и в качестве области I_β можно взять интервал, симметричный относительно 0, в который величина ξ, распределенная по Стьюденту, попадает с вероятностью β, т.е. $I_\beta = [-t_{n+m-2;\beta}, +t_{n+m-2;\beta}]$, где $P(|\xi| < t_{n+m-2;\beta}) = \beta$.

Таким образом, если нам заданы две выборки и уровень значимо-

сти α, мы вычисляем значение статистики t и ищем по α, n и m в табл. 6 (в ней содержатся критические значения распределения Стьюдента) значение $t_{n+m-2;\alpha}$. Если выполняется

$$\frac{|\bar{x}-\bar{y}|}{\sqrt{nS_x^2+mS_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}} < t_{n+m-2;\alpha},$$

то мы принимаем гипотезу о том что $\mu_x = \mu_y$. И отвергаем гипотезу $\mu_x = \mu_y$, если это неравенство не выполняется, так как произошло событие из дополнительной области, вероятность которой α. На рис. 4.1 заштрихованная площадь (вероятность попасть в область) равна $\beta = \int_{-t_\beta}^{+t_\beta} f_t(x)dx$, на рис. 4.2 заштрихована площадь $\alpha = 1-\beta$ области, где гипотеза не принимается.

Если оказалось, что $\bar{x} < \bar{y}$, можно проверять гипотезу о том, что $\mu_x = \mu_y$, когда альтернативной гипотезой является $\mu_x < \mu_y$. В этом случае строится "односторонняя" область, попадание в которую дает основание принять основную гипотезу. А именно, в табл. 6 в *нижней* строке отыскивается $\alpha = 1 - \beta$, где β – заданный уровень доверия, в строке с нужным числом степеней свободы находим границу интервала $t_{n+m-2;\alpha}$. Далее, если выполняется:

$$\frac{\bar{x}-\bar{y}}{\sqrt{nS_x^2+mS_y^2}}\sqrt{\frac{nm(n+m-2)}{n+m}} < -t_{n+m-2;\alpha},$$

то первая гипотеза неверна и принимается, что $\mu_x < \mu_y$.

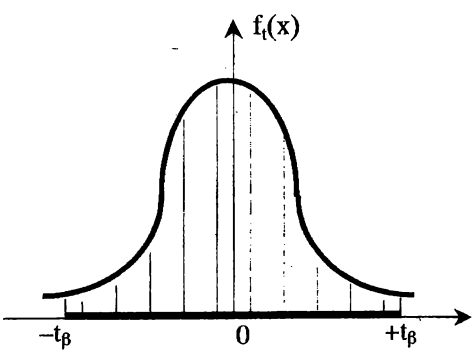


Рис. 4.1

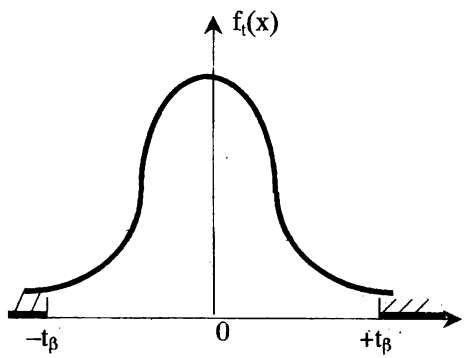


Рис. 4.2

Можно проверять гипотезу о том, что $\mu_x = \mu_y$, когда альтернативной гипотезой является $\mu_x > \mu_y$. В этом случае также строится "одностороняя" область, попадание в которую дает основание принять первую гипотезу (рис. 4.3). А именно, гипотеза о том, что $\mu_x \neq \mu_y$ не принимается, а принимается гипотеза $\mu_x > \mu_y$ тогда, когда

$$\frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} > t_{n+m-2;\alpha}.$$

С помощью нижней строки табл. 6 распределения Стьюдента (см. файл материалов) мы решали уравнения:

$$P\{\delta < -t_{n+m-2;\alpha}\} = \alpha, \text{ и } P\{\delta > t_{n+m-2;\alpha}\} = \alpha,$$

где α – уровень значимости.

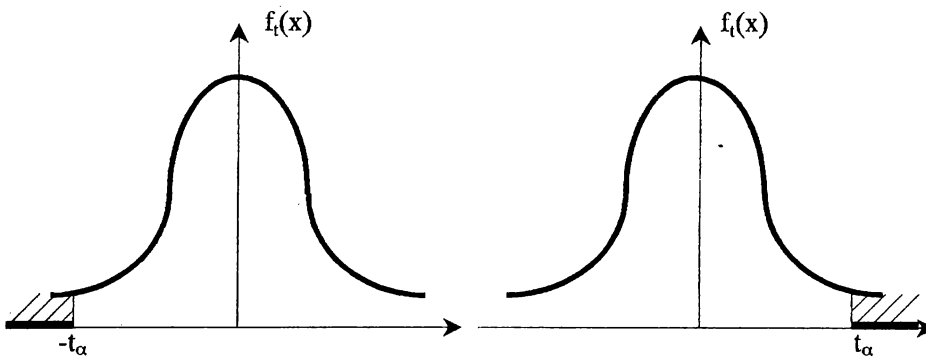


Рис. 4.3

Замечание. Было бы корректно сначала проверить гипотезу о равенстве дисперсий с помощью их выборочных оценок. Во второй части нашего руководства мы научимся делать такую проверку. Использовать значение статистики T можно только, если прошла гипотеза о равенстве дисперсий. Если дисперсии σ_x^2 и σ_y^2 неизвестны и не предполагается, что они равны, статистика T также имеет распределение Стьюдента. Но соответствующее ему число степеней свободы определяется приближенно и более сложным образом.

Итак, перечислим критерии, по которым проверяется статистическая гипотеза о том, что средние значения двух генеральных совокупностей, имеющих одинаковые дисперсии, совпадают ($\mu_x = \mu_y$) на уровне

значимости α. Они выведены из формул для двустороннего и одностороннего доверительного интервала для уровня доверия $\beta = 1-\alpha$.

Вычисляем по выборке значение статистики t:

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где

$$s^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{m}(y_i - \bar{y})^2\right] = \frac{1}{n+m-2}(nS_x^2 + mS_y^2).$$

1. Критическая область для односторонней проверки гипотезы, что средние значения двух генеральных совокупностей совпадают ($\mu_x = \mu_y$) по сравнению с альтернативой $\mu_x \neq \mu_y$ на уровне значимости α определяется неравенством:

$$|t| > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в верхней строке).

2. Критическая область для односторонней проверки гипотезы, что средние значения двух генеральных совокупностей совпадают ($\mu_x = \mu_y$) по сравнению с альтернативой $\mu_x > \mu_y$ на уровне значимости α определяется неравенством:

$$t > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

2. Критическая область для односторонней проверки гипотезы, что средние значения двух генеральных совокупностей совпадают ($\mu_x = \mu_y$) по сравнению с альтернативой $\mu_x < \mu_y$ на уровне значимости α определяется неравенством:

$$t < -t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

Если вычисленное значение статистики t попадает в критическую область, то основная гипотеза отвергается. Вероятность попадания в эту область равна принятому уровню значимости α. В этом случае принимается альтернативная гипотеза.

Пример 4.2. Результаты исследования двух сортов резины на покрышках (в баллах) приведены в таблице:

Номер покрышки	1	2	3	4
Износ для сорта А	32	40	36	35
Износ для сорта В	25	28	27	26

Сделать проверку гипотезы о том, что резина сорта А больше изнашивается, чем резина сорта В.

Решение.

$$\bar{x} = \frac{32 + 40 + 36 + 35}{4} = 35{,}75$$

$$\bar{y} = \frac{25 + 28 + 27 + 26}{4} = 26{,}5$$

$$nS_x^2 = 32^2 + 40^2 + 36^2 + 35^2 - 4 \cdot 35{,}75 \cdot 35{,}75 = 32{,}75;$$

$$nS_y^2 = 5;$$

$$t = \frac{9{,}25}{\sqrt{37{,}75}} \sqrt{\frac{4 \cdot 4 \cdot 6}{8}} = 5{,}2$$

Статистика распределена по Стьюденту с шестью степенями свободы. Значение t = 5,2 выходит далеко за пределы интервала, имеющего уровень доверия 0,99 (этот интервал (–3,14; 3,14)). Следовательно, гипотеза о том, что оба сорта резины изнашиваются одинаково, не проходит. А проходит альтернативная ей "одностороняя" гипотеза о том, что резина сорта А изнашивается сильнее.

Заметим, что сделанный вывод "заметен на глаз" и без расчетов. Такой результат получится во всех случаях, когда вычисленное по испытаниям значение статистики t > 4. При близости значения t к числам 2 или 3 вывод может включать в себя уровень доверия к нему как параметр (вспомните правила двух и трех σ). Если уровень доверия, с которым сделан вывод, не устраивает, надо позаботиться об улучшении опыта, например, увеличить число испытаний. Покажем это на примере.

Пример 4.3. Сравниваются две марки бензина А и В. На 11 машинах одинаковой мощности при разовом пробеге по кольцевому шоссе испытан бензин марок А и В. При испытании бензина марки В одна машина в пути вышла из строя и для нее данные по бензину отсутствуют. Расход бензина в пересчете на 100 км пути задан таблицей:

I	1	2	3	4	5	6	7	8	9	10	11	
x_i	10,51	11,86	10,5	9,1	9,21	10,74	10,75	10,8	11,3	11,8	10,9	N=11
y_i	13,23	13,0	11,5	10,4	11,8	11,6	10,64	12,3	11,1	11,6	—	m=10

Дисперсия расхода марок бензина А и В неизвестна и предполагается одинаковой. Можно ли с уровнем доверия 0,95 принять, что истинные средние расходы этих видов бензина одинаковы?

1. Вычисляем по выборкам:

$$\bar{x} = \frac{1}{11}\sum_{i=1}^{11} x_i = \frac{117{,}47}{11} = 10{,}68 \quad \bar{y} = \sum_{j=1}^{10} y_j = \frac{117{,}17}{10} = 11{,}2$$

$$Q = \sum_{i=1}^{11}(x_i - \bar{x})^2 + \sum_{j=1}^{10}(y_j - \bar{y})^2 = 14{,}8$$

$$t = \frac{10{,}68 - 11{,}72}{\sqrt{\frac{14{,}8}{19}} \cdot \sqrt{\frac{1}{11} + \frac{1}{10}}} = -1{,}04 \cdot 2{,}59 = -2{,}7$$

2. Находим из табл. 6 критических значений распределения Стьюдента $t_{19;0,95} = 2{,}1$.

3. Так как $|t| < t_{19;0,95}$ не выполняется, гипотезу о том, что средние значения норм расхода бензина марок А и В на 100 км пути совпадают, принять не можем, расхождение между $\bar{x}$ и $\bar{y}$ не объясняется только естественным разбросом данных.

4. Заметим, что если бы разброс Q оказался бы много больше, например, вдвое, то знаменатель увеличился в 1,41 раза и изменился бы и наш вывод – при таком большом разбросе расхождение между эмпирическими средними уже объяснялось бы естественным разбросом данных, а не расхождением теоретических средних.

5. Проверяем гипотезу о том, что $\mu_x < \mu_y$. С помощью нижней строки табл. 6 распределения Стьюдента решаем уравнение:

$$\int_{-\infty}^{-t} f_{19}(x)dx = 1 - \int_{-\infty}^{t} f_{19}(x)dx = 0{,}05 \quad \Rightarrow \quad -t = -1{,}8.$$

Так как $-2{,}7 < -1{,}8$, то гипотеза $\mu_x < \mu_y$ с уровнем доверия 0,95 может быть принята. По таблице видно, что эта гипотеза заслуживает доверия, которое можно оценить даже выше – в 99%.

Вывод. Средний расход бензина на 100 км для марки В больше, чем для марки А, с уровнем доверия 99%.

4.3. Ошибки первого и второго рода. Мощность критерия

Вывод о приемлемости основной гипотезы, ее непротиворечивости имеющимся данным не означает того, что доказана ее истинность. Принимая эту гипотезу, в некотором проценте случаев мы ошибемся. При принятии решения об истинности гипотезы возможны четыре случая:

Таблица 4.1

Гипотеза H_0	Принимается	Отвергается
Верна	Правильное решение	Ошибка 1-ого рода
Неверна	Ошибка 2-ого рода	Правильное решение

Ошибку α, когда отбрасывается основная гипотеза, хотя она истинна, называют **ошибкой первого рода;** в отличие от **ошибки второго рода** β, *которую совершают, приняв основную гипотезу, когда она ложна.* **Мощностью критерия** называется вероятность $(1-\beta)$ не допустить ошибку 2-го рода, т.е. отвергнуть гипотезу H_0, когда она неверна (это вероятность попадания критерия в критическую область при условии, что верна конкурирующая гипотеза).

На рис. 4.4 показано, какие площади изображают ошибку первого рода α, ошибку второго рода β и мощность критерия $1-\beta$ для случая, когда выборка производится из нормального распределения с известным среднеквадратическим отклонением σ. Критическая область строится с помощью распределения вероятностей статистики $\bar{x}$ для уровня значимости α. Проверяется гипотеза о том, что генеральное среднее равно a_0, а конкурирующей гипотезой является предположение о том, что среднее значение увеличилось и в качестве его нового значения берется значение a_1.

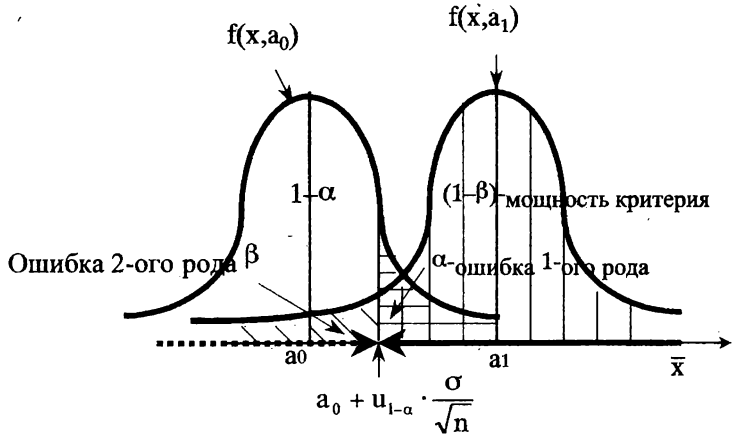


Рис. 4.4

На рис. 4.4 $f(x,a_0)$ и $f(x,a_1)$ – плотность распределения $\bar{x}$ при условии, что верна гипотеза; H_0 и H_1, соответственно, – нормальное распределение со среднеквадратическим отклонением $\frac{\sigma}{\sqrt{n}}$ и средним a_0 или a_1; через u_p обозначена квантиль стандартного нормального распределения, т.е. корень решения уравнения:

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{u_p} e^{-\frac{x^2}{2}}dx = P.$$

Так как f(x) нормальное распределение для вычисления этих площадей, используются таблицы квантилей нормального распределения. Ошибка 2-го рода β, в этом случае, вычисляется по a_0, a_1, σ и n.

Решение о выборе α и β зависит от конкретной задачи. Например, если отвергнуто правильное решение о "продолжении строительства", то эта ошибка (ошибка 1-го рода) повлечет материальный ущерб, если же принять решение о продолжении строительства, несмотря на опасность обвала, то это может повлечь и человеческие жертвы (ошибка 2-го рода). Можно привести примеры, когда ошибка 1-го рода повлечет более тяжкие последствия, чем ошибка 2-го рода.

4.4. Число испытаний при проверке гипотезы

Если мы хотим обеспечить, чтобы и ошибка первого рода не превосходила α, и неверность проверяемой гипотезы вскрывалась с вероятностью не меньшей, чем некоторое $1-\gamma$ (т.е. накладываем ограничение не только на α, но и требуем, чтобы выполнялось $\beta < \gamma$), может оказаться, что объем выборки должен быть увеличен.

Если $a_1 - u_{1-\gamma} \cdot \frac{\sigma}{\sqrt{n}} < a_0 + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ и значение статистики $\bar{x}$ попадает в интервал $a_1 - u_{1-\gamma} \cdot \frac{\sigma}{\sqrt{n}} < \bar{x} < a_0 + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}$ (рис. 4.5), то сделать выбор между гипотезами, ограничив ошибки 1-го и 2-го рода одновременно, нельзя. Для того, чтобы обеспечить требования к обеим ошибкам, n должно быть не меньше, чем то, при котором:

$$a_1 - u_{1-\gamma} \cdot \frac{\sigma}{\sqrt{n}} = a_0 + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}.$$

Следовательно, если мы хотим обеспечить, чтобы и ошибка 1-го рода не превосходила α и неверность проверяемой гипотезы вскрывалась с вероятностью не меньшей, чем некоторое $1-\gamma$, объем выборки должен быть не меньше, чем:

$\sqrt{n} \geq (u_{1-\alpha} + u_{1-\gamma}) \frac{\sigma}{|a_1 - a_0|}$ – для односторонней проверки и

$\sqrt{n} \geq (u_{1-\frac{1}{2}\alpha} + u_{1-\gamma}) \frac{\sigma}{\Delta}$ – для двусторонней проверки (Δ – отклонение второго среднего от первого)

Только при таком числе испытаний мы можем быть уверены в том, что, если верна гипотеза H_0, то мы ее отбросим с вероятностью, не меньшей, чем α; а если верна гипотеза H_1, то ее отбросить мы можем с вероятностью, не большей β.

Для того чтобы обеспечить требования к ошибкам 1-го и 2-го рода за минимальное число испытаний, можно применить процедуру последовательного статистического анализа.

При применении этого метода необходимое число наблюдений не фиксируется заранее, а определяется в процессе эксперимента.

Суть этого метода состоит в том, что область значений в n-мерном пространстве значений выборки делится на три части: критическую для

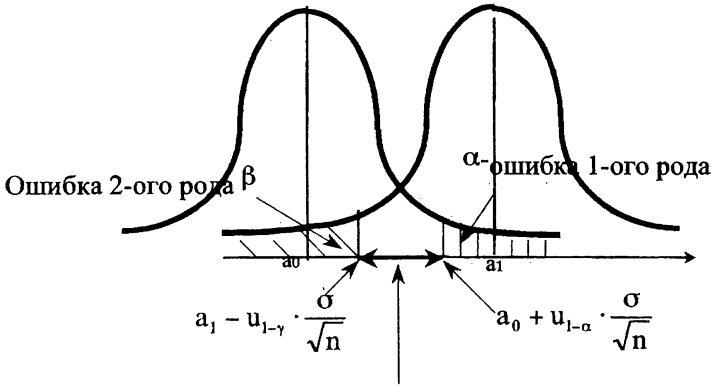


Рис. 4.5

гипотезы H_0 (ее вероятность при условии, что верна гипотеза H_0 равна α), критическую для гипотезы H_1 (ее вероятность при условии, что верна гипотеза H_1 равна β) и область неопределенности. Эксперимент продолжается, пока выборочные значения не попадут в одну из критических областей – или критическую для гипотезы H_0, или критическую для гипотезы H_1.

Недостатком метода последовательного анализа является необходимость на каждом шаге заново вычислять значения оценок. Но есть оценки, для которых это сделать нетрудно.

Например, обозначим через m номер шага.

$$X_m = \frac{1}{m}(x_1 + \ldots + x_m),$$

$$S_m^2 = \frac{1}{m}[(x_1 - X_m)^2 + \ldots + (x_1 - X_m)^2].$$

Пересчет оценок от шага к шагу можно провести по формулам:

$$X_m = \frac{1}{m}[X_{m-1}(m-1) + x_m],$$

$$S_m^2 = \frac{1}{m}[S_{m-1}^2(m-1) + (X_{m-1} - X_m)^2 + (x_m - X_m)^2]$$

Метод последовательного анализа построен на совсем иной теоретической основе, чем те методы, которые мы до сих пор рассматривали. Этот иной подход к решению задач основан на изучении функции правдоподобия.

5. МЕТОДЫ, ОСНОВАННЫЕ НА СВОЙСТВАХ ФУНКЦИИ ПРАВДОПОДОБИЯ

5.1. Функция правдоподобия

Функцией правдоподобия называется плотность вероятности (в дискретной модели просто вероятность) совместного появления результатов выборки x_1, x_2,...,x_n.

Таким образом, для распределения, зависящего от параметра Θ функция правдоподобия $L(x_1,x_2,...x_n,\Theta)$ задается по формуле:

$$L(x_1,x_2,...x_n,\Theta) = f(x_1,\Theta)f(x_2,\Theta)...f(x_n,\Theta) = \prod_{i=1}^{n} f(x_i,\Theta),$$

где $f(x,\Theta)$ – плотность распределения в случае непрерывной модели и вероятность значения x в дискретной модели.

5.2. Метод максимального правдоподобия для получения оценки неизвестного параметра Θ

Р.Фишером для получения оценки неизвестного параметра Θ был в 1912 г. предложен метод, обладающий оптимальными свойствами, который называется методом максимального правдоподобия.

Согласно **методу максимального правдоподобия** в качестве оценки неизвестного параметра Θ принимается такое значение Θ_n, при котором плотность вероятности (в дискретной модели просто вероятность) совместного появления результатов выборки x_1, x_2,...x_n. максимальна, и, следовательно, функция правдоподобия достигает максимума.

В точке, в которой значение функции L максимально, ее производная по параметру Θ обращается в ноль. Чаще всего ищется не максимум функции L, а максимум ее логарифма, так как максимум этих функций достигается при одном и том же значении Θ. Следовательно, для нахождения оценки требуется решить уравнение (или, если параметров несколько, систему уравнений) правдоподобия:

$$\frac{d\ln L}{d\Theta} = 0 \text{ или } \frac{1}{L}\frac{dL}{d\Theta} = 0.$$

Основной недостаток метода максимального правдоподобия – трудность вычисления оценок, связанных с решением уравнения правдоподобия. Кроме того, необходимо знать тип анализируемого закона распределения $f(x,\Theta)$, что во многих случаях оказывается практически нереальным. Достоинством метода является то, что при достаточно общих условиях оценки максимального правдоподобия являются состоятельными, асимптотически эффективными и имеют асимптотически нормальное распределение.

Пример 5.1. Методом максимального правдоподобия найти оценку для вероятности p наступления события A в схеме Бернулли (n раз проводятся независимые испытания, при которых возможны только два исхода – с вероятностью p происходит событие A, с вероятностью $q = 1-p$ событие A не происходит). В результате n испытаний событие A произошло m раз.

Выписываем функцию правдоподобия:

$$L = C_n^m p^m (1-p)^{n-m},$$

$$\ln L = C + m \cdot \ln p + (n-m) \cdot \ln(1-p).$$

$$\frac{d \ln L}{dp} = \frac{m}{p} - \frac{n-m}{1-p}, \quad \text{откуда} \quad \tilde{p} = \frac{m}{n}.$$

Таким образом, оценкой метода максимального правдоподобия вероятности p события A является относительная частота m/n этого события.

Пример 5.2. Выборка производится из нормального распределения $N(a,\sigma)$.

$$L = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_1-a)^2}{2\sigma^2}} \cdot \ldots \cdot \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_n-a)^2}{2\sigma^2}}$$

Следовательно, $\ln L$ равен

$$\ln L = C + \frac{1}{2\sigma^2} \sum_{i=1}^{n} (x_i - a)^2.$$

Приравняв нулю производные по a и по σ, можно показать, что для его параметров a и σ^2 оценками наибольшего правдоподобия будут те же $\bar{x}$ и S^2, которые мы получили в качестве оценок для этих параметров методом моментов.

5.3. Функции правдоподобия в задачах проверки гипотез

Рассмотрим простейшую задачу проверки гипотезы, когда конкурирующими являются две простые гипотезы: H_0 – значение неизвестного параметра распределения равно a_0, и альтернативная ей H_1 – значение неизвестного параметра распределения равно a_1. На основе статистического эксперимента надо принять решение о том, какая из гипотез правильная.

Составим **отношение вероятностей** L_n (отношение функций правдоподобия для конкурирующих гипотез) для n испытаний:

$$L_n = \frac{\prod_{i=1}^{n} f(x_i, a_1)}{\prod_{i=1}^{n} f(x_i, a_0)}.$$

Задачу о конкурирующих гипотезах можно решать, изучая поведение отношения вероятностей. Согласно теореме Неймана-Пирсона, там, где оно больше некоторого порога ($L_n > C$ или $\ln L_n > \ln C$), следует предпочесть гипотезу H_1, в противном случае – H_0. Порог C надо принять таким, чтобы обеспечить уверенность в том, что, если верна гипотеза H_0, то мы ее отбросим с вероятностью не большей, чем α. Симметричным рассуждением (поменяв местами гипотезы), находим порог для значений отношения вероятностей такой, что если верна гипотеза H_1, то ее мы отбросим с вероятностью, не большей чем β.

Пример 5.3. Производятся испытания с величиной ξ, которая распределена нормально с известной дисперсией σ^2. Относительно математического ожидания a имеются две гипотезы: H_0 состоит в том, что $a = a_0$, и H_1 – состоит в том, что $a=a_1$. Вычислим $\ln L_n$.

$$\ln L_n = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[(x_i - a_1)^2 - (x_i - a_0)^2] =$$

$$= \frac{(a_1 - a_0)}{\sigma^2}\sum_{i=1}^{n} x_i - \frac{n}{2\sigma^2}(a_1^2 - a_0^2) \quad .$$

Таким образом, вместо порога функции правдоподобия, обеспечивающего ошибку α, можно искать порог для статистики $X_n = \frac{1}{n}(x_1 + \ldots + x_n)$. Порог ищется по распределению вероятностей. Эта статистика имеет

нормальное распределение. Тем же способом найдем и порог, обеспечивающий требуемую величину ошибки 2-го рода. Если верна гипотеза H_0, то $X_n \sim N(a_0, \frac{\sigma}{\sqrt{n}})$. Если H_1, то $X_n \sim N(a_1, \frac{\sigma}{\sqrt{n}})$. Задача свелась к рассмотренному выше примеру. Надо вычислить необходимое число испытаний n по формуле:

$$\sqrt{n} = (u_{1-\alpha} + u_{1-\beta}) \frac{\sigma}{|a_1 - a_0|}$$

и пороговое соотношение для X_n из уравнения:

$$\frac{(X_n - a_0)\sqrt{n}}{\sigma} = u_{1-\alpha},$$

с учетом вычисленного значения n. Это дает:

$$X_n \geq a_0 + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = a_0 + u_{1-\alpha} \cdot \frac{a_1 - a_0}{u_{1-\alpha} + u_{1-\beta}} = a_0 \frac{u_{1-\beta}}{u_{1-\alpha} + u_{1-\beta}} + a_1 \frac{u_{1-\alpha}}{u_{1-\alpha} + u_{1-\beta}}$$

Таким образом, в терминах отношения правдоподобия процедура проверки гипотезы о среднем, когда конкурируют две простые гипотезы для значений среднего a_0 и a_1 при генеральном среднеквадратическом отклонении σ для заданных ошибок 1-го и 2-го рода α и β, выглядит следующим образом:

1. Вычисляется число испытаний n по формуле:

$$\sqrt{n} = (u_{1-\alpha} + u_{1-\beta}) \frac{\sigma}{|a_1 - a_0|}.$$

2. Для величины X_n вычисляется порог, определяющий критическую область критерия. Критическая область, при попадании в которую гипотеза H_0 отвергается, имеет вид:

$$X_n \geq a_0 + u_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} = a_0 + u_{1-\alpha} \cdot \frac{a_1 - a_0}{u_{1-\alpha} + u_{1-\beta}} = a_0 \frac{u_{1-\beta}}{u_{1-\alpha} + u_{1-\beta}} + a_1 \frac{u_{1-\alpha}}{u_{1-\alpha} + u_{1-\beta}}$$

На рис. 5.1 изображены плотности распределения статистики X_n и критические области для случая $\alpha = \beta$. При $\alpha = \beta$ критерий выглядит очень просто. Число испытаний находится по формуле:

$$\sqrt{n} = 2u_{1-\alpha} \frac{\sigma}{a_1 - a_0}.$$

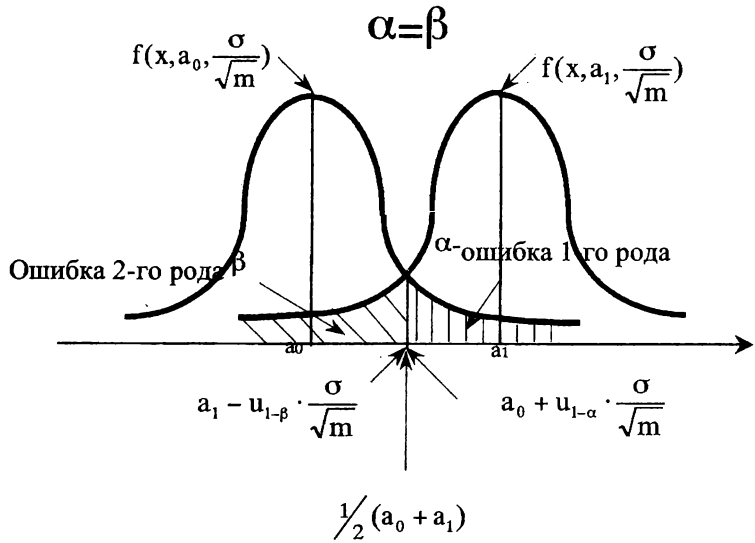


Рис. 5.1

Если $X_n > \frac{1}{2}(a_0 + a_1)$ – H_0 отклоняется.

Если $X_n < \frac{1}{2}(a_0 + a_1)$ – H_0 принимается.

(Какая плотность больше, та и считается верной).

5.4. Проверка гипотез методом последовательного анализа

Когда проведение каждого испытания стоит очень дорого, для того чтобы при проверке статистических гипотез минимизировать число испытаний, рекомендуется применять последовательные критерии.

Последовательные критерии впервые были предложены А.Вальдом в 1946 г. Как уже было сказано, при применении этого метода (последовательного анализа) необходимое число наблюдений не фиксируется заранее, как мы это делали, например, в только что рассмотренной задаче, а определяется в процессе эксперимента; область значений в n-мерном пространстве делится на три части: критическую для гипотезы H_0 (ее вероятность при условии, что верна гипотеза H_0 – α), критическую для гипотезы H_1 (ее вероятность при условии, что верна гипотеза H_1 – β) и область неопределенности. Эксперимент продолжается, пока выборочные значения не попадут в одну из критических областей – или критическую для гипотезы H_0, или критическую для гипотезы H_1.

В самом общем виде по этому методу последовательность наблюдений проводится по следующей схеме.

1. Вычисляются границы критерия:

$$A = (1-\beta)/\alpha,\ B = \beta/(1-\alpha),$$

где α – вероятность отклонить основную гипотезу, когда она верна,

β – вероятность принять основную гипотезу, когда она неверна.

2. На каждом шаге вычисляется отношение вероятностей L_m:

$$L_m = \frac{\prod_{i=1}^{m} f(x_i, a_1)}{\prod_{i=1}^{m} f(x_i, a_0)}.$$

В этом отношении a_0 – значение параметра для основной гипотезы, a_1 – для альтернативной, m – текущее число наблюдений или номер шага, $f(x, \Theta)$ – плотность вероятности для непрерывной модели и вероятность принять значение x для дискретной модели. Обычно удобнее вычислять $\ln L_m$.

3. Если $\ln L_m \le \ln B$, то принимается основная гипотеза.

Если $\ln L_m \ge \ln A$, то основная гипотеза отклоняется. Если не выполняется ни одно из неравенств, эксперимент продолжается.

Доказана теорема о том, что с вероятностью 1 эта процедура заканчивается за конечное число шагов.

Организация эксперимента по **методу последовательного статистического анализа** позволяет уменьшить число испытаний n *в среднем вдвое* (а при $\alpha = \beta$ даже в 4 раза) по сравнению с оптимальным методом с фиксированным числом наблюдений.

Пример 5.4. Рассмотрим применение последовательного анализа в условиях предыдущей задачи, когда с помощью эксперимента мы хотим решить вопрос о среднем значении нормального распределения, имеющего среднеквадратическое отклонение σ, когда заданы ошибки 1-го и 2-го рода α и β. Процедура последовательного анализа будет выглядеть следующим образом.

Вычислим $A = (1-\beta)/\alpha$ и $B = \beta/(1-\alpha)$. На каждом шаге эксперимента вычисляется значение $\ln L_m$.

$$\ln L_m = -\frac{1}{2\sigma^2}\sum_{i=1}^{m}[(x_i - a_1)^2 - (x_i - a_0)^2] =$$

$$= \frac{(a_1 - a_0)}{\sigma^2}\sum_{i=1}^{m} x_i - \frac{m}{2\sigma^2}(a_1^2 - a_0^2) = \frac{(a_1 - a_0)m}{\sigma^2}\left(X_m - \frac{a_1 + a_0}{2}\right).$$

Решение о гипотезе принимается в соответствии с пунктом 3 решающего правила:

$$X_m \le \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)} \ln\frac{\beta}{1-\alpha}$$ – принимается H_0.

$$X_m \ge \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)} \ln\frac{1-\beta}{\alpha}$$ – принимается H_1.

Эксперимент продолжается, если:

$$\frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)} \ln\frac{\beta}{1-\alpha} < X_m < \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)} \ln\frac{1-\beta}{\alpha}$$

Для упрощения расчетов рассмотрим случай:
$a_0 = 0$; $a_1 = 1$; $\sigma = 1$; $\alpha = \beta = 0{,}025$;
$\ln\beta/(1-\alpha) = \ln 0{,}026 = -3{,}6$;
$\ln(1-\beta)/\alpha = \ln 39 = 3{,}6$;
H_0 отклоняется, если $X_m > 1/2(a_0+a_1) + 1/m \cdot 3{,}6$;
H_1 принимается, если $X_m < 1/2(a_0+a_1) - 1/m \cdot 3{,}6$.
Эксперимент продолжается, если:
$1/2(a_0+a_1) - 1/m \cdot 3{,}6 < X_m < 1/2(a_0+a_1) + 1/m \cdot 3{,}6$.

Так как $u_{1-0{,}025} = 2$, то при n = 16 (см. выше), если эксперимент еще не закончится, можно воспользоваться критерием, предлагаемым методом с фиксированным числом наблюдений:

Если $X_m > 1/2(a_0+a_1)$ – H_0 отклоняется.

Если $X_m < 1/2(a_0+a_1)$ – H_0 принимается.

Схематически для $\alpha = \beta$ это можно изобразить следующим образом (рис. 5.2):

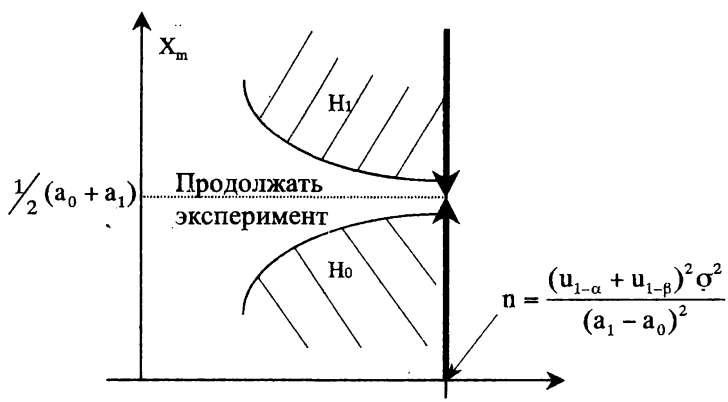


Рис. 5.2

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

1. Составьте логическую схему базы знаний по прилагаемому файлу материалов и перечень основных зависимостей и формул.

2. Решите следующие задачи, пользуясь таблицами файла материалов:

2.1. Из Таблицы 1 чисел выборки из равномерного распределения на отрезке [0,100] возьмите подряд 100 чисел, начиная с номера 4N, где N – ваш порядковый номер в списке группы (дойдя до конца таблицы, перейдите в ее начало). Возьмите в качестве интервалов группировки интервалы (0, 20), (20, 40)...(80, 100) и напишите таблицу эмпирического распределения для этих интервалов. По этой таблице постройте гистограмму и полигон, сосчитайте эмпирические среднее, дисперсию ($\bar{x}$,S^2), эмпирическое среднеквадратическое отклонение. Выпишите теоретические значения для этих величин и сравните их с эмпирическими.

2.2. Из Таблицы 2 чисел выборки из нормального распределения N(0,1) возьмите подряд 100 чисел, начиная с номера 4N, где N – ваш порядковый номер в списке группы (дойдя до конца таблицы, перейдите в ее начало). Возьмите в качестве интервалов группировки интервалы (-3,-2), (-2,-1)...(2,3) и напишите таблицу эмпирического распределения для этих интервалов. По этой таблице постройте гистограмму и полигон, сосчитайте эмпирические среднее и дисперсию ($\bar{x}$,S^2), эмпирическое среднеквадратическое отклонение. Выпишите теоретические значения для этих величин и сравните их с эмпирическими.

2.3. В условиях предыдущей задачи построить 95%-ый доверительный интервал для математического ожидания генеральной совокупности, при условии, что дисперсия генерального распределения известна и равна 1. Попало ли оцениваемое значение в доверительный интервал?

2.4. Задание то же, что в п. 2.3, но считать, что дисперсия генерального распределения неизвестна. Попало ли оцениваемое значение в доверительный интервал? Сильно ли различаются интервалы, построенные в этой и предыдущей задаче?

2.5. Производитель стальных канатов долгое время обеспечивал прочность каната на разрыв μ = 55000 кг при стандартном отклонении σ = 500 кг. После усовершенствования процесса изготовления, производитель стал утверждать, что прочность каната на разрыв возросла. При испытании выборки из n = 50 канатов получено, что средняя выборочная прочность составляет 55250 кг. Заказчик решил проверить гипотезу H_0: μ = 55000 при уровне значимости 0,05 (так как он сомневается в увеличении μ). Пройдет ли эта гипотеза?

2.6. Для двух нормальных независимых величин ξ и η: $\xi \sim N(\mu_\xi, \sigma)$ и $\eta \sim N(\mu_\eta, \sigma)$ с одинаковыми дисперсиями получены выборки объема $n_\xi = 42$ и $n_\eta = 20$, для которых сосчитано: $\overline{\xi} = 64$, $S_\xi^2 = 16$, $\overline{\eta} = 62$, $S_\eta^2 = 25$. При уровне значимости α = 0,05 проверяется гипотеза H_0: $\mu_\xi = \mu_\eta$ о равенстве генеральных средних (альтернативная гипотеза H_1: $\mu_\xi \neq \mu_\eta$). Чему равно опытное значение статистики T, применяемой для проверки гипотезы H_0?

2.7. Чему равна в задаче 2.6 область принятия гипотезы H_0? Можно ли принять гипотезу H_0?

2.8. Если $\overline{\xi} = 64$, $S_\xi^2 = 16$, $\overline{\eta} = 61$, $S_\eta^2 = 25$, то каково будет решение?

2.9. Из проверяемых на всхожесть 8000 зерен случайным образом отобрано 1000. Среди них оказалось (84+N)% недоброкачественных (N – ваш номер в списке). Найти доверительную вероятность того, что процент таких зерен в генеральной совокупности отличается от процента их в выборке не более, чем на 2% (по абсолютной величине). Рассмотреть случаи повторной и бесповторной выборки.

ТРЕНИНГ УМЕНИЙ

1. Пример выполнения упражнений тренинга на умение № 1

Задание

Построить гистограмму и полигон по заданной таблице:

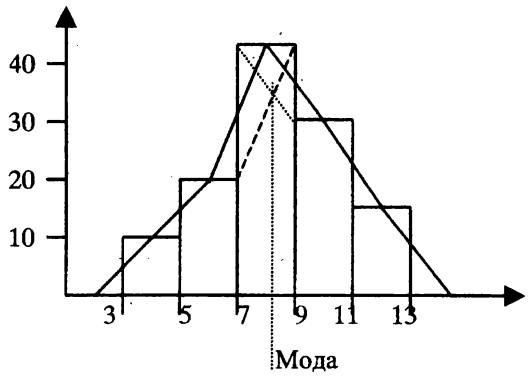
Распределение семей по размеру жилой площади, приходящейся на одного человека (цифры условные)

№	Площадь, приходящаяся на одного человека	Число семей с данным размером площади
1	3-5	10
2	5-7	20
3	7-9	40
4	9-11	30
5	11-13	15
	Всего	115

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Упорядочить заданные значения по возрастанию, сосчитать их количество	$n = 115$ Упорядочивать значения не требуется, так как задана интервальная таблица.
2.	Сгруппировать значения, если надо; сосчитать число значений, попавших в интервалы разбиения; вычислить эмпирические частоты, составить таблицу эмпирического распределения	Таблица частот появления значения Значения 4 6 8 10 12 Кол-во m_i 10 20 40 30 15 Таблица эмпирического распределения x_i 4 6 8 10 12 m_i/n 0,087 0,174 0,348 0,261 0,130

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
3.	По таблице эмпирического распределения нарисовать гистограмму и полигон, найти медиану	 **Гистограмма и полигон** Медиана 8,375 (делит площадь гистограммы пополам)

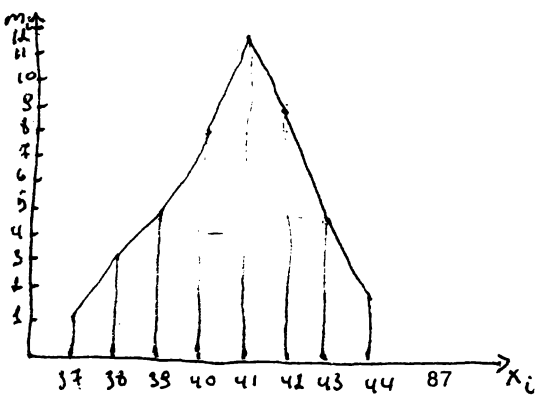
Решите самостоятельно следующие задачи:

1.1. Построить дискретный вариационный ряд и начертить полигон для следующего распределения размеров 45 пар мужской обуви, проданных в магазине за день:

39 41 40 42 41 40 42 44 40 43 42 41 43 39 42 41 42 39 41 37 43 41 38 43 42 41 40 41 38 44 40 39 41 40 42 40 41 42 40 43 38 39 41 41 42.

Найти моду и медиану.

x_i	37	38	39	40	41	42	43	44
m_i	1	3	5	8	12	8	5	2



1.2. Наблюдения за жирностью молока у 50 коров дали следующие результаты (в %):

3,86 4,06 3,67 3,97 3,76 3,61 3,96 4,04 3,84 3,94 3,98 3,57 3,87 4,07 3,99 3,69 3,76 3,71 3,94 3,82 4,16 3,76 4,00 3,46 4,08 3,88 4,01, 3,93 3,71 3,81 4,02 4,17 3,72 4,09 3,78 4,02 3,73 3,52 3,89 3,92 4,18 4,26 4,03 4,14 3,72 4,33 3,82 4,03 3,62 3,91

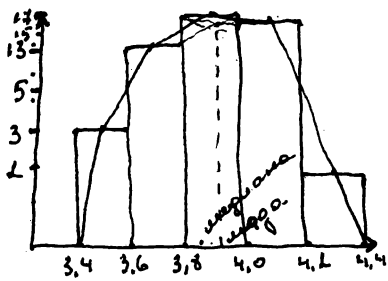
Построить по этим данным интервальный вариационный ряд с равными интервалами (например, первый интервал 3,40-3,60, второй – 3,60-3,80 и т.д.) и изобразить его графически – нарисовать гистограмму и полигон. Найти моду и медиану.

x_i	3,4 - 3,6	3,6 - 3,8	3,8 - 4,0	4,0 - 4,2	4,2 - 4,4
m_i	3	13	17	15	2

$\sum m_i = 50.$

$x_{ср} = \frac{1}{50}(3,5 + 3,7 + 3,8 + 4,1 + 4,3) =$

$\approx 3,8$



2. Пример выполнения упражнений тренинга на умение № 2

Задание 1

Для случайно отобранных семи рабочих стаж работы оказался равным: 10, 3, 5, 12, 11, 7, 9.

Чему равен для них средний стаж и чему равен разброс (среднеквадратическое отклонение)?

Решение

Заполните таблицу, подобрав в каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Выписать заданные значения, объем выборки и нужную формулу для получения точечной оценки	Задана выборка: x_1 x_2 x_3 x_4 x_5 x_6 x_7 10, 3, 5, 12, 11, 7, 9. n = 7; $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ - формула для среднего $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \left(\frac{1}{n}\right)\sum_{i=1}^{n} x_i^2 - \bar{x}^2$ – формула для дисперсии $\sigma = \sqrt{S^2}$ - среднеквадратическое отклонение.
2.	Сосчитать значение оценки	$\bar{x} = \frac{10+3+5+12+11+7+9}{7} = 8{,}14$ года $S^2 = \frac{1}{7}(10^2+3^2+5^2+12^2+11^2+7^2+9^2) -$ $-(8{,}14)^2 = 75{,}57 - 66{,}26 = 9{,}31$ $\sigma = \sqrt{9{,}31} = 3{,}05$ года

Задание 2

При обследовании надоя коров случайным образом отобрали 307 коров, данные по ним сгруппировали и составили таблицу:

Надои	3000-3400	3400-3800	3800-4200	4200-4600	4600-5000
Число коров	43	71	102	64	27

Найти выборочное среднее, дисперсию и среднеквадратическое отклонение.

Решение

Заполните таблицу, подобрав в каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Выписать заданные значения, объем выборки и нужную формулу для получения точечной оценки	Составляем таблицу числа наблюдения значений x_i 3200 3600 4000 4400 4800 m_i 43 71 102 64 27 $n = 307$ $\overline{x} = \sum_{j=1}^{k} x_j \frac{m_j}{n} = \frac{1}{n}\sum_{j=1}^{k} x_j m_j$ – выборочное среднее $S^2 = \sum_{j=1}^{k} x_j^2 \frac{m_j}{n} - \overline{x}^2 = \sum_{j=1}^{r} (x_j - \overline{x})^2 \frac{m_j}{n}$ – выборочная дисперсия $\sigma = \sqrt{S^2}$ - выборочное среднеквадратическое отклонение.
2.	Сосчитать значение оценки	$\overline{x} = 1/307(3200 + 3600 + 4000 + 4400 + 4800) = 3949;$ $S^2 = 1/307(3200^2 \cdot 43 + 3600^2 \cdot 71 + 4000^2 \cdot 102 +$ $+ 4400^2 \cdot 64 + 4800^2 \cdot 27) - (3949)^2 = 215170;$ $\sigma = \sqrt{215170} = 463{,}86$ литров

Решите самостоятельно следующую задачу:

2.1. Построить таблицу дискретного вариационного ряда, начертить полигон распределения 60 абитуриентов по числу баллов, полученных ими на приемных экзаменах. Найти эмпирические моду, медиану, среднее значение и среднеквадратическое отклонение:

20 19 22 24 21 18 23 17 20 16 15 23 21 24 21 18 23 21 19 20 24 21 20 18 17 22 20 16 22 18 20 17 21 17 19 20 20 21 18 22 23 21 25 22 20 19 21 24 23 21 19 22 21 19 20 23 22 25 21 21

3. Пример выполнения упражнений тренинга на умение № 3

Задание

С целью определения времени, затрачиваемого на обработку детали, взяты выборочно 100 рабочих крупного завода. Результаты обследования приведены в таблице:

Время обработки в минутах	3,6-4,2	4,2-4,8	4,8-5,4	5,4-6,0	6,0-6,6
Число рабочих	14	33	35	12	6

Требуется найти выборочное среднее, дисперсию, среднее квадратическое отклонение и границы, в которых с надежностью 0,95 заключено среднее время обработки детали всеми рабочими завода.

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Сосчитать выборочное среднее, выборочное среднеквадратическое отклонение	Составляем таблицу числа наблюдения значений x_i 3,9 4,5 5,1 5,7 6,3 m_i 14 33 35 12 6 n = 100 $\bar{x} = \sum_{j=1}^{k} x_j \frac{m_j}{n} = 1/n \sum_{j=1}^{k} x_j m_j$ – выборочное среднее,

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
	(если не известно истинное), выписать нужную формулу доверительного интервала	в данном случае $\bar{x} = 4{,}88$; $S^2 = \sum_{j=1}^{k} x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^{r} (x_j - \bar{x})^2 \frac{m_j}{n}$ – выборочная дисперсия, в данном случае $S^2 = 0{,}38$; $\sigma = \sqrt{S^2}$ – выборочное среднеквадратическое отклонение, в данном случае $\sigma = 0{,}62$. Доверительный интервал для истинного среднего времени обработки детали: $\bar{x} - t_{n-1;\beta} \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_{n-1;\beta} \frac{S}{\sqrt{n-1}}$. $\beta = 0{,}95$, $t_{99;0{,}95}$ находим в таблице 6 для распределения Стьюдента (при таком n можно воспользоваться таблицей 5 для нормального распределения).
2.	Пользуясь таблицами 5 или 6, вычислить границы требуемого в задании интервала; выписать полученный доверительный интервал	$t_{99;0{,}95} = 2$ $4{,}88 - \frac{2 \cdot 0{,}64}{\sqrt{99}} < \mu < 4{,}88 + \frac{2 \cdot 0{,}64}{\sqrt{99}}$ $4{,}76 < \mu < 5$

Решите самостоятельно следующие задачи:

3.1. Для определения зольности угля месторождения взято 200 проб. В результате лабораторных исследований установлена средняя зольность угля в выработке 17% при среднем квадратическом отклонении 3%. С вероятностью 0,95 определите пределы, в которых находится средняя зольность угля месторождения μ. Постройте односторонние пределы для того же уровня доверия (не меньше, чем ... и не больше, чем ...)

$n = 200$
$\bar{x} = 0{,}17 = 17\%$
$\sigma = 0{,}03 = 3\%$
$\beta = 0{,}95$
$\mu = ?$

$\bar{x} - t_{n-1;\beta} \frac{S}{\sqrt{n-1}} < \mu < \bar{x} + t_{n-1;\beta} \frac{S}{\sqrt{n-1}}$

$t_{n-1;\beta} = t_{199;0{,}95} = 1{,}96 \approx 2$

$\sigma = \sqrt{S^2} = S = 0{,}03$

$0{,}17 - 2 \frac{0{,}03}{\sqrt{199}} < \mu < 0{,}17 + 2 \frac{0{,}03}{\sqrt{199}}$

$0{,}166 < \mu < 0{,}174$ ← довер-й интервал.

√ 3.2. Из партии подшипников было случайным образом отобрано 8 деталей и сделаны замеры на точность обработки (в мкм): 216,54; 216,53; 216,51; 216,56; 216,57; 216,55; 216,52; 216,54. Найти несмещенные оценки математического ожидания и дисперсии замеров. Определить доверительные интервалы для математического ожидания с надежностью 0,95. *Указание:* при расчетах вычесть из заданных значений 216 и использовать свойства математического ожидания и дисперсии: $M(X+C) = M_x + C$; $D(X+C) = D_x$.

4. Пример выполнения упражнений тренинга на умение № 4

Задание 1

Выборочная проверка показала, что из 100 изделий 87 удовлетворяют стандарту. Мы хотим быть уверены на 95%, что не ошибаемся в оценке процента нестандартных изделий. В каких пределах он находится? Каков должен быть объем выборки, чтобы оценить процент брака с точностью до 0,01?

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Вычислить оценку $\tilde{p}$ для p	$n = 100$; $\tilde{p} = 0{,}13$
2.	Найти доверительный интервал для p	$\tilde{p} - 1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + 1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$ (применили формулу для повторной выборки – с возвратом). Подставив n и p, получаем $0{,}06 < p < 0{,}2$
	В случае, когда требуется, проверить гипотезу; сформулировать вывод из эксперимента, провести вычисления с доверительным интервалом	С вероятностью 0,95 выполняется: $\lvert p - \tilde{p}\rvert \le 1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$. Требуемая точность 0,01; следовательно, $1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{N}} = 0{,}01;$ $N = \frac{\tilde{p}(1-\tilde{p})\cdot 1{,}96^2}{0{,}01^2} = 38416\tilde{p}(1-\tilde{p}) =$ $= 38416 \cdot 0{,}87 \cdot 0{,}13 = 4345$

Задание 2

Партия изделий считается годной к выпуску, если брак в ней не превышает 3%. Из партии в 2000 изделий было отобрано и проверено 400. При этом бракованных оказалось 6. Какова вероятность того. что вся партия удовлетворяет техническим условиям и может быть принята?

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Вычислить оценку $\tilde{p}$ для p	$N = 2000$; $n = 400$; $\tilde{p} = 6/400 = 0{,}015$
2.	Найти доверительный интервал для p	$\tilde{p} - k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\sqrt{1-\frac{n}{N}} \le p \le \tilde{p} + k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\sqrt{1-\frac{n}{N}}$ применили формулу для бесповторной выборки (без возврата)
3.	В случае, когда требуется, проверить гипотезу; сформулировать вывод из эксперимента, провести вычисления с доверительным интервалом	Следовательно, должно выполняться: $\tilde{p} + k_\beta\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\sqrt{1-\frac{n}{N}} \le 0{,}03$. Подставляем наши данные: $k_\beta = \frac{20 \cdot 0{,}015}{\sqrt{0{,}015 \cdot 0{,}985 \cdot 0{,}8}} = 2{,}75$. Соответствующее такому k_β значение вероятности β находим в Таблице 4 функции Лапласа Ф(x): $\beta = 0{,}994$. С вероятностью 0,994 выполняется: $p < 0{,}03$

Решите самостоятельно следующие задачи:

4.1. Выборочно обследовали качество кирпича. Из 1600 проб в 32 случаях кирпич оказался бракованным. Требуется определить, в каких пределах заключается доля брака для всей продукции, если результат необходимо гарантировать с вероятностью 0,954.

$$\tilde{p} - 1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + 1{,}96\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}$$

$n = 1600$

$m = 32$

$\tilde{p} = \frac{m}{n} = \frac{32}{1600} = 0{,}02$

$0{,}0068$

$$0{,}02 - 1{,}96\sqrt{\frac{0{,}02(1-0{,}02)}{1600}} \le 0{,}02 \le 0{,}02 + 1{,}96\sqrt{\frac{0{,}02(1-0{,}02}{1600}} =$$

$$= 0{,}0132 \le p \le 0{,}0268.$$

4.2. В выборке объемом 500 единиц, произведенной для определения процента всхожести зерна р установлена относительная частота доброкачественных зерен k/n = 0,94. Найти, с какой вероятностью может быть принят в этом случае искомый процент всхожести, если допустимая погрешность в его определении равна ±2%.

5. Пример выполнения упражнений тренинга на умение № 5

Задание

Провели обследование однотипных изделий, произведенных двумя заводами (по 40 изделий на каждом заводе). Оценки вычислялись в некоторых единицах, затем по ним для каждого завода были сосчитаны статистические показатели – среднее значение оценки и среднеквадратическое отклонение. Результаты приведены в таблице:

	Завод № 1	Завод № 2
Средний балл	71	76
Стандартное отклонение	5	6

Проверить при уровне значимости 0,05 гипотезу о том, что изделия завода № 2 лучшего качества, чем изделия завода № 1.

Решение

Заполните таблицу, подобрав каждому алгоритму конкретное содержание.

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
1.	Выписать из условия задачи данные о выборке. Сосчитать оценки для среднего и дисперсии	$\bar{x} = 71; \bar{y} = 76; S_x = 5; S_y = 6;$ $n = 40; m = 40; \alpha = 0{,}05.$
2.	Сформулировать проверяемую гипотезу в вероятностных терминах. Выписать формулу статистики, вычисляемой по выборке. Выписать число степеней свободы N для распределения статистики. Подставить в формулу статистики данные выборки	Проверяется гипотеза о том, что $\mu_x = \mu_y$, когда альтернативной гипотезой является гипотеза $\mu_x < \mu_y$ (вариант 2с). Пользуемся статистикой $T = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$, $N = 78.$ Вычисляем значение статистики: $T = \frac{71 - 76}{\sqrt{40 \cdot 5^2 + 40 \cdot 6^2}} \sqrt{\frac{40 \cdot 40 \cdot 78}{80}} = -4$

№ п/п	Алгоритм	Конкретное соответствие данной ситуации предложенному алгоритму
3.	Выписать критическую область и с помощью таблиц найти границы критической области для статистики, с помощью которой будет проверяться гипотеза	Критическая область для проверки гипотезы: $T < -t_{N;\alpha}$ ($t_{N;\alpha}$ отыскивается по таблице 6 критических значений распределения Стьюдента, α в нижней строке). Если вычисленное значение попадет внутрь интервала $[-\infty, -1{,}66]$, то гипотеза H_0 не пройдет, а пройдет интересующая нас гипотеза H_1.
4.	Сформулировать вывод, требуемый в задаче	$-4 < -1{,}66$, следовательно, значение статистики попало в критическую область и проходит не основная, а альтернативная гипотеза о том, что изделия завода № 1 хуже изделий завода № 2, (это практически достоверно)

Решите самостоятельно следующие задачи:

5.1. На некотором поле выбрали 100 участков земли: 50 участков засеяли одним сортом ячменя, 50 – другим. На первых 50 участках получили урожай в среднем 60 ц/га со стандартным отклонением 3ц/га, на вторых 50 участках средний урожай оказался равным 65 ц/га со стандартным отклонением 3,5 ц/га. Будет ли средний урожай этого сорта ячменя значимо превосходить средний урожай ячменя первого сорта? Принять $\alpha = 0{,}05$.

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{n S_x^2 + m S_y^2}} \cdot \sqrt{\frac{n m (n + m - 2}{n + m}}.$$

$\bar{x} = 60$
$\bar{y} = 65$
$S_x = 3$
$S_y = 3{,}5$
$n = 50$
$m = 50$
$N = n + m - 2 = 98$
$\alpha = 0{,}05$

H_0; μ

$$T = \frac{60 - 65}{\sqrt{50 \cdot 3^2 + 50 \cdot 3{,}5^2}} \cdot \sqrt{\frac{50 \cdot 50 (\cdot 98}{50 + 50}} =$$

$$= \frac{-5}{\sout{45{,}4} \; 32{,}6} \cdot 48{,}5 = -7{,}6$$

$T = -7{,}6$. знач-е статистики

$T \leq t_{N;\alpha}$ $t_{98; 0{,}05} = 1{,}66$

Так как нер-во верно
[illegible] гип-за отверг.

2 - сорт лучше чем первый.

5.2. Средняя месячная выработка для выборки из 50 рабочих одного завода составляет 110 изделий при среднеквадратическом отклонении 14 изделий, а для выборки из 40 рабочих другого завода равна 105 изделий при среднеквадратическом отклонении 15 изделий. Выше ли средняя выработка на первом заводе, чем на втором? Уровень значимости α принять равным 0,01.

5.3. Из генеральной совокупности извлечена выборка объема n = 50.

Варианта x_i	-2	1	2	3	4	5
Частота m_i	10	5	10	10	10	5

Оценить с надежностью 0,95 математическое ожидание a нормально распределенного признака генеральной совокупности по выборочной средней при помощи доверительного интервала. (*Указание:* при $n > 20$ распределение Стьюдента практически совпадает с нормальным распределением). Пройдет ли при уровне значимости $\alpha = 0{,}05$ гипотеза о том, что генеральное среднее равно 3? Дайте объяснение, почему гипотеза проходит, или не проходит.

ФАЙЛ МАТЕРИАЛОВ

Приложение 1

ГРЕЧЕСКИЙ АЛФАВИТ

A	α	альфа	N	ν	ню
B	β	бета	Ξ	ξ	кси
Γ	γ	гамма	O	o	омикрон
Δ	δ	дельта	Π	π	пи
E	ε	эпсилон	P	ρ	ро
Z	ζ	дзета	Σ	σ	сигма
H	η	эта	T	τ	тау
Θ	θ	тета	Y	υ	ипсилон
I	ι	йота	Φ	φ	фи
K	κ	каппа	X	χ	хи
Λ	λ	ламбда	Ψ	ψ	пси
M	μ	мю	Ω	ω	омега

СПИСОК ФОРМУЛ

ТЕОРИЯ ВЕРОЯТНОСТЕЙ

1. Функция распределения F(x) и плотность f(x) и их свойства

$$0 \le F(x) \le 1.$$

$$F(x_1) \le F(x_2)$$

$$\lim_{x \to -\infty} F(x) = 0, \quad \lim_{x \to \infty} F(x) = 1$$

$$\int_{-\infty}^{\infty} f(x)dx = F(\infty) = 1$$

$$F(x) = p(\xi < x) = \int_{-\infty}^{x} f(t)dt$$

$$f(x) = F'(x)$$

$$p(x_1 \le \xi < x_2) = \int_{x_1}^{x_2} f(t)dt$$

$$p(\xi > x) = \int_{x}^{\infty} f(t)dt = 1 - F(x)$$

2. Математическое ожидание и его свойства

Для дискретного распределения:

$$M(\xi) = a = \sum_{i=1}^{n} x_i \cdot p_i$$

Для непрерывного распределения:

$$M(\xi) = a = \int_{-\infty}^{\infty} x \cdot f(x)dx$$

$$M(C) = C$$

$$M(C\cdot\xi) = C\cdot M(\xi)$$

$$M\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} M(\xi_i)$$

$$M(\xi+C) = M(\xi)+C$$

3. Дисперсия, среднеквадратическое отклонение и их свойства

$$D(\xi) = \sigma^2 = M((\xi - a)^2)$$

$$D(\xi) = \sum_{i=1}^{n}(x_i - a)^2 \cdot p_i = \sum_{i=1}^{n} x_i^{\,2} \cdot p_i - a^2 \text{ (дискретный случай)}$$

$$D(\xi) = \int_{-\infty}^{\infty}(x - a)^2 \cdot f(x)dx = \int_{-\infty}^{\infty} x^2 \cdot f(x)dx - a^2 \text{ (непрерывный случай).}$$

$$D(C) = 0$$

$$D(C\cdot\xi) = C^2\cdot D(\xi)$$

$$D\left(\sum_{i=1}^{n}\xi_i\right) = \sum_{i=1}^{n} D(\xi_i) \text{ (для независимых } \xi_i)$$

$$D(\xi+c) = D(\xi),\ D(\xi-\eta) = D(\xi) + D(\eta)$$

$$\sigma(\xi) = \sqrt{D(\xi)}$$

$$\sigma(C) = 0;\ \sigma(C\cdot\xi) = |C|\cdot\sigma(\xi)$$

4. Примеры распределений и значения их числовых характеристик

4.1. Испытание с двумя исходами, биномиальное распределение

Пусть в результате испытания с вероятностью p происходит событие A (случайная величина ζ – индикатор наступления A приняла значе-

ние 1), а с вероятностью q = 1–p противоположное ему событие $\bar{A}$ (ζ приняла значение 0). Это распределение задают таблицей:

x_i	0	1
p_i	q	p

$$M(\zeta) = 0\cdot q + 1\cdot p = p$$

$$D(\zeta) = 0\cdot q + 1^2\cdot p - p^2 = p\cdot(1-p) = pq$$

Вероятность, что число успехов ξ, полученных при n независимых испытаниях, проводящихся над такой случайной величиной (схема Бернулли), примет значение m, задается **формулой Бернулли**:

$$p_n(m) = C_n^m \cdot p^m \cdot q^{n-m} = \frac{n\cdot(n-1)\cdot\ldots\cdot(n-m+1)}{1\cdot 2\cdot\ldots\cdot m} p^m q^{n-m}$$

$$M(\xi) = M\left(\sum_{k=1}^{n}\zeta_k\right) = \sum_{k=1}^{n} M(\zeta_k) = np,\ D(\xi) = D\left(\sum \zeta_k\right) = \sum_{k=1}^{n} D(\zeta_k) = npq$$

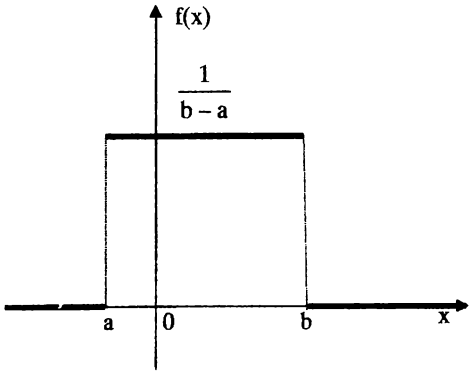
4.2. Распределение Пуассона

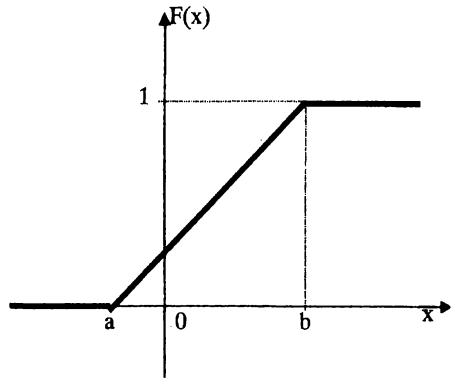
Случайная величина, которая принимает значение m с вероятностью $P_m(\lambda) = \frac{\lambda^m e^{-\lambda}}{m!}$, где m=0,1,2,..., а λ – положительная постоянная величина, называется **распределенной по Пуассону с параметром λ.**

$$M(\xi) = \sum_{m=0}^{\infty} \frac{m\cdot\lambda^m}{m!} e^{-\lambda} = \lambda \sum_{m=0}^{\infty} \frac{\lambda^m}{m!} e^{-\lambda} = \lambda,\ D(\xi) = \lambda$$

4.3. Равномерное распределение

$$f(x) = \begin{cases} \frac{1}{b-a}, & x\in[a,b] \\ 0, & x\notin[a,b] \end{cases};\quad F(x) = \begin{cases} 0, & x\in(-\infty,a) \\ \frac{x-a}{b-a}, & x\in[a,b] \\ 1, & x\in(b,\infty) \end{cases};\quad \begin{matrix} M(\xi) = \frac{b+a}{2} \\ D(\xi) = \frac{(b-a)^2}{12} \end{matrix}$$

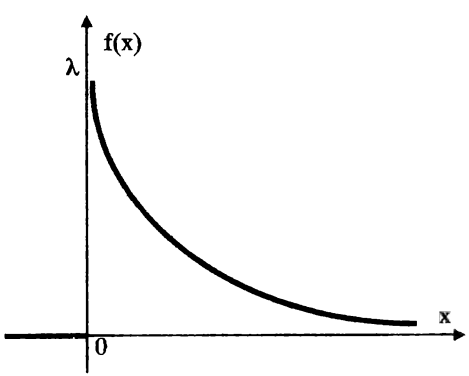


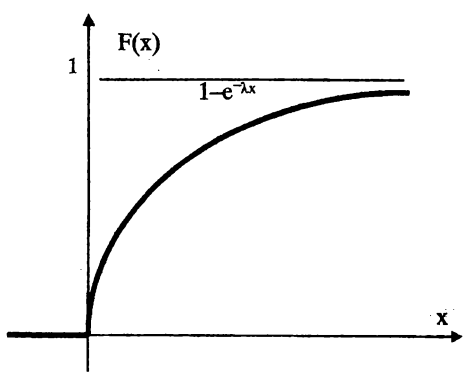


4.4. Показательное распределение

Непрерывная случайная величина имеет показательное распределение, если её плотность вероятности имеет вид

$$f(x)=\begin{cases}\lambda\cdot e^{-\lambda\cdot x}, & x\in[0,\infty);\ M(\xi)=\lambda\int\limits_0^{\infty}xe^{-\lambda x}dx=\dfrac{1}{\lambda};\\ 0, & x\in(-\infty,0);\ D(\xi)=\lambda\int\limits_0^{\infty}(x-\dfrac{1}{\lambda})^2e^{-\lambda x}dx=\dfrac{1}{\lambda^2};\end{cases}$$





4.5. Нормальное распределение

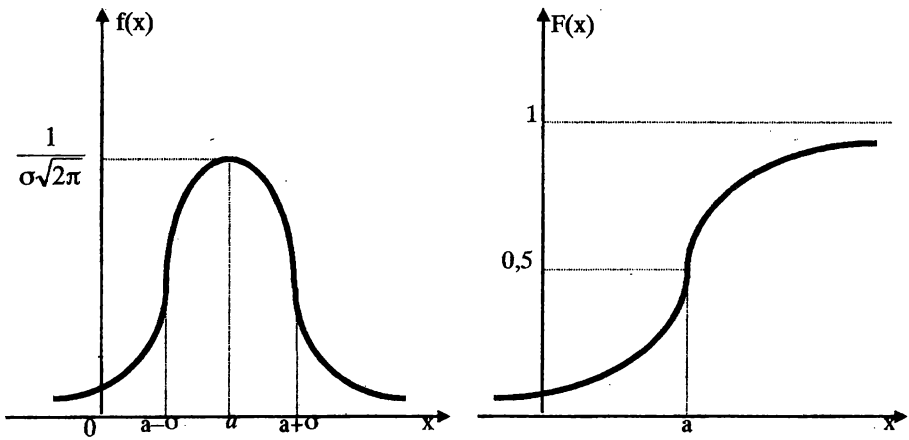
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-a)^2}{2\sigma^2}}$$

$$M(\xi) = a,\ D(\xi) = \sigma^2,\ s(\xi) = \sigma$$

Параметрами, определяющими нормальное распределение N(a,σ) являются a – математическое ожидание и σ – среднеквадратическое отклонение.

ξ~N(0,1) называют стандартной нормальной. Ее плотность вероятности и функция распределения задаются формулами:

$$f_\xi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}};\quad F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt$$ (Таблица 3 файла материалов)



$$\Phi(x) = \frac{2}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{t^2}{2}} dt$$ – функция Лапласа (Таблица 4)

Функция распределения стандартного нормального закона F(x) связана с функцией Лапласа Φ(x) соотношением:

$$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} dt = \int_{-\infty}^{0} + \int_{0}^{x} = 0.5 + 0.5 \cdot \Phi(x) = 0.5[1 + \Phi(x)];$$

Для $\xi \sim N(0,1)$:

$$P(A \le \xi \le B) = \frac{1}{\sqrt{2\pi}}\int_A^B e^{-\frac{t^2}{2}}dt = \frac{1}{\sqrt{2\pi}}\int_A^0 e^{-\frac{t^2}{2}}dt + \frac{1}{\sqrt{2\pi}}\int_0^B e^{-\frac{t^2}{2}}dt =$$

$$= \frac{1}{2}[\Phi(B) - \Phi(A)]$$

Для $\xi \sim N(a,\sigma)$:

$$P(A \le \xi \le B) = \frac{1}{\sigma\sqrt{2\pi}}\int_A^B e^{-\frac{(t-a)^2}{2\sigma^2}}dt = \frac{1}{\sqrt{2\pi}}\int_{\frac{A-a}{\sigma}}^{\frac{B-a}{\sigma}} e^{-\frac{y^2}{2}}dy =$$

$$= \frac{1}{2}\left[\Phi\left(\frac{B-a}{\sigma}\right) - \Phi\left(\frac{A-a}{\sigma}\right)\right]$$

Правило трёх σ: $p(|\xi - a| \le 3\sigma) = \Phi(3) = 0,9973$

Правило двух σ: $p(|\xi - a| \le 2\sigma) = \Phi(2) = 0,9544$

$$p\{|\xi - a| < k_\beta\sigma\} = \Phi(k_\beta) = \beta$$

Вероятности попадания в полуинтервал для $\xi \sim N(a,\sigma)$:

$$P(\xi \le B) = \frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^B e^{-\frac{(t-a)^2}{2\sigma^2}}dt = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\frac{B-a}{\sigma}} e^{-\frac{y^2}{2}}dy =$$

$$= F\left(\frac{B-a}{\sigma}\right) = 0,5\left[1 + \Phi\left(\frac{B-a}{\sigma}\right)\right]$$

$$P(\xi > B) = \frac{1}{\sigma\sqrt{2\pi}}\int_B^\infty e^{-\frac{(t-a)^2}{2\sigma^2}}dt = \frac{1}{\sqrt{2\pi}}\int_{\frac{B-a}{\sigma}}^\infty e^{-\frac{y^2}{2}}dy =$$

$$= 1 - F\left(\frac{B-a}{\sigma}\right) = 0,5\left[1 - \Phi\left(\frac{B-a}{\sigma}\right)\right]$$

$$P\{\xi < a - u_\beta\sigma\} = P\{\xi > a + u_\beta\sigma\} =$$

$$= \frac{1}{2}[1 - \Phi(u_\beta)] = F(-u_\beta) = 1 - F(u_\beta) = 1 - \beta = \alpha$$

В частности:

$$P\{\xi < a - 1{,}65\sigma\} = P\{\xi > a + 1{,}65\sigma\} = 0.05$$

$$P\{\xi < a - 2\sigma\} = P\{\xi > a + 2\sigma\} = 0.025$$

$$P\{\xi > a - 1{,}65\sigma\} = P\{\xi < a + 1{,}65\sigma\} = 0.95$$

$$P\{\xi > a - 2\sigma\} = P\{\xi < a + 2\sigma\} = 0.975$$

5. Нормировка

Если $M(\xi) = a$, $D(\xi) = \sigma^2$, то $M\left(\frac{\xi - a}{\sigma}\right) = 0$; $D\left(\frac{\xi - a}{\sigma}\right) = 1$.

МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

1. Выборка

$$\sum_{i=1}^{k} m_i = n \;;\; \tilde{p}_i = \frac{m_i}{n} \;;\; \sum_{i=1}^{k} \tilde{p}_i = 1$$

Для дискретного вариационного ряда медиана d:

$$d = \begin{cases} \dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, & \text{если } n \text{ четно} \\ x_{\frac{n+1}{2}}, & \text{если } n \text{ нечетно} \end{cases}$$

Размах вариационного ряда – расстояние $x_{max} - x_{min}$ между крайними членами вариационного ряда

Группировка состоит в том, что область на оси x, куда попали значения $x_1, x_2, \ldots, x_n$, разбивают на интервалы $l_1, l_2, \ldots, l_k$ и подсчитывают частоту попадания значений величины в каждый интервал. Самый простой способ группировки – округление данных

Согласно формуле Серджеса рекомендуемое число интервалов:

$$k = 1 + 3{,}322 \lg n,$$

а величину интервала h можно вычислить по формуле:

$$h = \frac{x_{max} - x_{min}}{1 + 3{,}322 \lg n},$$

где $x_{max} - x_{min}$ разность между наибольшим и наименьшим значением в выборке (ее размах). За начало первого интервала рекомендуется брать величину:

$$x_{нач} = x_{min} - 0{,}5h.$$

Кроме того, необходимо следить, чтобы не было интервалов, в которые попало меньше пяти значений.

2. Теоретические и эмпирические моменты

Начальный момент *l*-ого порядка a_l:

$a_l = \sum x_i^l p_i$ для дискретного распределения и

$a_l = \int x^l f(x)\,dx$ для непрерывного распределения.

***l*-ый центральный момент b_l:**

$b_l = \sum (x_i - \mu)^l p_i$ для дискретного распределения и

$b_l = \int (x - \mu)^l f(x)\,dx$ для непрерывного распределения,

где $\mu = a_1$ – математическое ожидание распределения

Вариационный ряд без повторов

***l*-ый начальный эмпирический момент**: $a_l = \frac{1}{n}\sum_{i=1}^{n} x_i^l$

***l*-ый центральный эмпирический момент**: $b_l = \frac{1}{n}\sum_{j=1}^{n} (x_j - \bar{x})^l$

Вариационный ряд, заданный таблицей

$$a_l = \sum_{j=1}^{k} x_j^l \frac{m_j}{n} = \sum_{j=1}^{k} x_j^l \tilde{p}_j$$

$$b_l = \sum_{j=1}^{k} \left(x_j - \bar{x}\right)^l \frac{m_j}{n} = \sum_{j=1}^{k} \left(x_j - \bar{x}\right)^l \tilde{p}_j$$

3. Выборочное (эмпирическое) среднее $\bar{x}$

Выборка задана вариационным рядом: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Выборка задана таблицей: $\bar{x} = \frac{1}{n}\sum_{i=1}^{k} x_j m_j = \sum_{j=1}^{k} x_j \frac{m_j}{n} = \sum_{j=1}^{k} x_j \tilde{p}_j$.

Свойства те же, что у математического ожидания.

Распределение выборочного среднего: $\bar{x} \approx N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.

4. Выборочная (эмпирическая) дисперсия S^2

Выборка задана вариационным рядом:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2$$

Выборка задана таблицей:

$$S^2 = \frac{1}{n}\sum_{j=1}^{k} x_j^2 m_j - \bar{x}^2 = \sum_{j=1}^{k} x_j^2 \frac{m_j}{n} - \bar{x}^2 = \sum_{j=1}^{k} x_j^2 \tilde{p}_j - \bar{x}^2 =$$

$$= \sum_{j=1}^{k} (x_j - \bar{x})^2 \frac{m_j}{n} = \sum_{j=1}^{k} (x_j - \bar{x})^2 \tilde{p}_j$$

Свойства те же, что у дисперсии.

Исправленная, несмещенная оценка для дисперсии:

$$s^2 = \frac{n}{n-1} S^2$$

5. Доверительные интервалы

Доверительный интервал с уровнем доверия β для математического ожидания a нормального распределения для случая, когда известно среднеквадратическое отклонение распределения σ:

Двусторонний: $\bar{x} - k_\beta \frac{\sigma}{\sqrt{n}} < a < \bar{x} + k_\beta \frac{\sigma}{\sqrt{n}}$ (табл. 5 β в 1-м столбце)

односторонний: $-\infty < a < \bar{x} + k_\beta \frac{\sigma}{\sqrt{n}}$ (табл. 5 $\alpha = 1-\beta$ в 3-м столбце)

и $\bar{x} - k_\beta \frac{\sigma}{\sqrt{n}} < a < \infty$ (Таблица 5 $\alpha = 1-\beta$ в 3-м столбце)

Доверительный интервал с уровнем доверия β для математического ожидания a нормального распределения для случая, когда среднеквадратическое отклонение распределения σ неизвестно:

Двусторонний: $\bar{x} - t_{n-1,\beta} \frac{s}{\sqrt{n}} < a < \bar{x} + t_{n-1,\beta} \frac{s}{\sqrt{n}}$

или

$$\bar{x} - t_{n-1,\beta} \frac{S}{\sqrt{n-1}} < a < \bar{x} + t_{n-1,\beta} \frac{S}{\sqrt{n-1}},$$

где S – корень из эмпирической дисперсии, а s – корень из исправленной эмпирической дисперсии (табл. 6 β в верхней строке)

односторонние: $-\infty < a < \bar{x} + t_{n-1,\beta} \frac{s}{\sqrt{n}}$ (табл. 6 β в нижней строке) и

$\bar{x} - t_{n-1,\beta} \frac{s}{\sqrt{n}} < a < \infty$ (таблица 6 β в нижней строке).

В этих формулах s^2 несмещенная оценка для дисперсии:

$$s^2 = \frac{n}{n-1} S^2.$$

6. Оценка требуемого объема выборки

Минимальный объем выборки n, обеспечивающий чтобы точность оценки, полученной по ней для a с надежностью β, не превосходила заданного значения ε (то есть $|\bar{x} - a| < \varepsilon$), когда среднеквадратическое отклонение известно, задается формулой:

$$n = \left(\frac{k_\beta \sigma}{\varepsilon} \right)^2$$

7. Доверительный интервал для вероятности успеха в схеме Бернулли

Основная формула – следствие интегральной теоремы Муавра-Лапласа, из которой выводятся любые соотношения между эмпирической частотой, генеральной частотой, n и вероятностью β:

$$P\left\{\left|\frac{m-np}{\sqrt{npq}}\right| < k_\beta\right\} = P\left\{|m-np| < k_\beta \cdot \sqrt{npq}\right\} =$$

$$= P\left\{\left|\frac{m}{n} - p\right| < k_\beta \sqrt{\frac{pq}{n}}\right\} \cong \Phi(k_\beta) = \beta.$$

Отсюда формула доверительного интервала для выборки с возвратом:

$$\tilde{p} - k_\beta \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} \le p \le \tilde{p} + k_\beta \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}, \text{ где } \tilde{p} = \frac{m}{n}.$$

Для выборки без возврата формулу надо подправить:

$$\tilde{p} - k_\beta \sqrt{\frac{m/n(1-m/n)}{n}} \sqrt{1-\frac{n}{N}} \le p \le \tilde{p} + k_\beta \sqrt{\frac{m/n(1-m/n)}{n}} \sqrt{1-\frac{n}{N}}.$$

Объем выборки n, обеспечивающий, чтобы точность оценки, полученной по ней для p с надежностью β, не превосходила заданного значения ε, т.е. $|\tilde{p} - p| < \varepsilon$, находится по формуле:

$$n = \frac{k_\beta^2}{\varepsilon^2} \tilde{p}(1-\tilde{p}).$$

8. Критерии проверки статистических гипотез о средних

Выведены из формул для двустороннего и одностороннего доверительного интервала для уровня доверия $\beta = 1-\alpha$.

8.1. Проверка гипотезы $\mu=\mu_0$ (уровень значимости равен α).

По выборке вычисляется значение статистики:

$$T = \frac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}}.$$

a) Гипотеза H_0: $\mu=\mu_0$, альтернативная H_1: $\mu\neq\mu_0$. Критическая область для проверки гипотезы:

$$|T| > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в верхней строке).

b) Гипотеза H_0: $\mu=\mu_0$, альтернативная H_1: $\mu>\mu_0$. Критическая область для проверки гипотезы:

$$T > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

c) Гипотеза H_0: $\mu=\mu_0$, альтернативная H_1: $\mu<\mu_0$. Критическая область для проверки гипотезы:

$$T < -t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

8.2. Проверка гипотезы $\mu_x=\mu_y$ (уровень значимости равен α). По выборке вычисляется значение статистики:

$$T = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = \frac{\bar{x} - \bar{y}}{\sqrt{nS_x^2 + mS_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}},$$

где

$$s^2 = \frac{1}{n+m-2}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2 + \sum_{j=1}^{m}(y_i - \bar{y})^2\right] = \frac{1}{n+m-2}(nS_x^2 + mS_y^2)$$

a) Гипотеза H_0: $\mu_x=\mu_y$ альтернативная H_1: $\mu_x\neq\mu_y$. Критическая область для проверки гипотезы:

$$|T| > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в верхней строке).

b) Гипотеза H_0: $\mu_x=\mu_y$, альтернативная H_1: $\mu_x>\mu_y$. Критическая область для проверки гипотезы:

$$T > t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

c) Гипотеза H_0: $\mu_x=\mu_y$, альтернативная H_1: $\mu_x<\mu_y$. Критическая область для проверки гипотезы:

$$T < -t_{n-1;\alpha}$$

($t_{n-1;\alpha}$ отыскивается по табл. 6 критических значений распределения Стьюдента, α в нижней строке).

Если вычисленное значение статистики T попадает в критическую область, то основная гипотеза H_0 отвергается, и принимается альтернативная гипотеза H_1. Вероятность попадания в эту область равна уровню значимости α.

9. Понятие квантили

Значения u_p, для которых выполняется $\int_{-\infty}^{u_p} f(x)dx = P$, называются **квантилями** распределения f(x).

10. Выбор между двумя простыми гипотезами с использованием отношения функций правдоподобия

Функция правдоподобия $L(x_1,x_2,...x_n,\Theta)$ задается по формуле:

$$L(x_1,x_2,...x_n,\Theta) = f(x_1,\Theta)f(x_2,\Theta)...f(x_n,\Theta) = \prod_{i=1}^{n} f(x_i,\Theta),$$

где $f(x,\Theta)$ – плотность распределения в случае непрерывной модели и вероятность значения x в дискретной модели.

Отношение вероятностей L_n (отношение функций правдоподобия для конкурирующих гипотез) для n испытаний:

$$L_n = \frac{\prod_{i=1}^{n} f(x_i,a_1)}{\prod_{i=1}^{n} f(x_i,a_0)}.$$

Если надо обеспечить, чтобы и ошибка 1-го рода не превосходила α и неверность проверяемой гипотезы вскрывалась с вероятностью не меньшей, чем некоторое $1-\gamma$, объем выборки должен быть не меньше, чем:

$$\sqrt{n} \geq (u_{1-\alpha} + u_{1-\gamma})\frac{\sigma}{|a_1 - a_0|}$$ – для односторонней проверки и

$$\sqrt{n} \geq (u_{1-\frac{1}{2}\alpha} + u_{1-\gamma})\frac{\sigma}{|a_1 - a_0|}$$ – для двусторонней проверки

Процедура проверки гипотезы о среднем, когда конкурируют две простые гипотезы для значений среднего a_0 и a_1 при генеральном среднеквадратическом отклонении σ для заданных ошибок 1-го и 2-го рода α и β.

А. Проверка с фиксированным числом испытаний.

1. Вычисляется число испытаний n по формуле:

$$\sqrt{n} = (u_{1-\alpha} + u_{1-\beta})\frac{\sigma}{|a_1 - a_0|}$$

2. Для величины X_n вычисляется порог, определяющий критическую область критерия. Критическая область, при попадании в которую гипотеза H_0 отвергается, имеет вид:

$$X_n \geq a_0 + u_{1-\alpha}\cdot\frac{\sigma}{\sqrt{n}} = a_0 + u_{1-\alpha}\cdot\frac{a_1 - a_0}{u_{1-\alpha} + u_{1-\beta}} = a_0\frac{u_{1-\beta}}{u_{1-\alpha} + u_{1-\beta}} + a_1\frac{u_{1-\alpha}}{u_{1-\alpha} + u_{1-\beta}}$$

В. Проверка той же гипотезы, проводимая по методу последовательного анализа.

Решение о гипотезе принимается в соответствии с правилом:

$$X_m \leq \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)}\ln\frac{\beta}{1-\alpha}$$ – принимается H_0;

$$X_m \geq \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)}\ln\frac{1-\beta}{\alpha}$$ – принимается H_1.

Эксперимент продолжается, если:

$$\frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)}\ln\frac{\beta}{1-\alpha} < X_m < \frac{a_0 + a_1}{2} + \frac{\sigma^2}{m(a_1 - a_0)}\ln\frac{1-\beta}{\alpha}.$$

В обеих процедурах $X_m = 1/m(x_1 + x_2 + \ldots + x_m)$. Для процедуры последовательного анализа полезна формула последовательного вычисления $X_m : X_m = 1/m[(m-1)\,X_{m-1} + x_m]$.

Приложение 3

Таблица 1

Последовательность случайных чисел, распределенных равномерно на отрезке [0,100]

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76	80	95	90	91	17
37	54	20	48	05	64	89	47	42	96	24	80	52	40	37	20	63	61	04	02
08	42	26	89	53	19	64	50	93	03	23	20	90	25	60	15	95	33	47	64
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65	88	67	67	43	97
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53	98	95	11	68	77
66	06	57	47	17	34	07	27	68	50	36	69	73	61	70	65	81	33	98	85
31	06	01	08	05	45	57	18	24	06	35	30	34	26	14	86	79	90	74	39
85	26	97	76	02	02	05	16	56	92	68	66	57	48	18	73	05	38	52	47
22	15	67	16	01	76	72	52	73	62	79	88	03	40	47	40	99	58	39	51
05	94	66	77	42	77	53	12	97	87	01	95	47	73	83	68	41	90	12	26

Таблица 2

Последовательность случайных чисел, имеющих распределение N(0,1)

0,414	0,011	0,666	-1,132	-0,410	-1,077	1,484	-0,340	0,789	-0,494	0,364
-1,237	-0,044	-0,111	-0,210	0,931	0,616	-0,377	-0,433	1,048	-0,037	0,759
0,609	-2,043	-2,290	0,404	-0,543	0,486	0,869	0,347	2,816	-0,464	-0,632
-1,614	0,372	-0,074	-0,916	1,314	-0,038	0,673	0,563	-0,107	0,131	-1,808
0,284	0,458	1,307	-1,625	-0,629	-0,504	-0,0056	-0,131	0,048	1,879	-1,016
0,360	-0,119	2,331	1,672	-1,053	0,840	0,246	-0,237	-1,312	1,603	-0,952
-0,566	1,600	0,465	1,951	0,110	0,251	0,116	-0,957	-0,190	1,479	-0,986
1,249	1,934	0,070	-1,358	-1,246	-0,959	-1,297	-0,722	0,925	0,783	-0,402
0,619	1,826	1,272	-0,945	0,494	0,050	-1,696	1,876	0,063	0,132	0,682
0,544	-0,417	-0,666	-0,104	-0,253	-2,543	-1,133	1,987	0,668	0,360	1,927
1,183	1,211	1,765	0,035	-0,359	0,193	-1,023	-0,222	-0,616	-0,060	-1,319
-0,785	-0,430	-0,298	0,248	-0,088	-1,379	0,295	-0,115	-0,621	-0,618	0,209
0,979	0,906	-0,096	-1,376	1,047	-0,872	-2,200	-1,384	1,425	-0,812	0,748
-1,095										

А. Квантили нормального распределения $\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{u_P} e^{-\frac{t^2}{2}}dt = P$

P	u_p	P	u_p	P	u_p	P	u_p	P	u_p
0,0	$-\infty$	0,2	-0,842	0,4	-0,253	0,6	2,253	0,8	0,842
0,01	-2,326	0,21	-0,806	0,41	-0,228	0,61	0,279	0,81	0,878
0,02	-2,054	0,22	-0,772	0,42	-0,202	0,62	0,305	0,82	0,915
0,03	-1,881	0,23	-0,739	0,43	-0,176	0,63	0,332	0,83	0,954
0,04	-1,751	0,24	-0,706	0,44	-0,151	0,64	0,358	0,84	0,994
0,05	-1,645	0,25	-0,674	0,45	-0,126	0,65	0,385	0,85	1,036
0,06	-1,555	0,26	-0,643	0,46	-0,10	0,66	0,412	0,86	1,080
0,07	-1,476	0,27	-0,613	0,47	-0,075	0,67	0,440	0,87	1,126
0,08	-1,405	0,28	-0,583	0,48	-0,50	0,68	0,468	0,88	1,175
0,09	-1,341	0,29	-0,553	0,49	-0,025	0,69	0,496	0,89	1,227
0,1	-1,282	0,3	-0,524	0,5	0	0,7	0,524	0,9	1,282
0,11	-1,227	0,31	-0,496	0,51	0,025	0,71	0,553	0,91	1,341
0,12	-1,175	0,32	-0,468	0,52	0,050	0,72	0,583	0,92	1,405
0,13	-1,126	0,33	-0,440	0,53	0,075	0,73	0,613	0,93	1,476
0,14	-1,080	0,34	-0,412	0,54	0,1	0,74	0,643	0,94	1,555
0,15	-1,036	0,35	-0,385	0,55	0,126	0,75	0,674	0,95	1,645
0,16	-0,994	0,36	-0,358	0,56	0,151	0,76	0,706	0,96	1,751
0,17	-0,954	0,37	-0,332	0,57	0,176	0,77	0,739	0,97	1,881
0,18	-0,915	0,38	-0,305	0,58	0,202	0,78	0,772	0,98	2,054
0,19	-0,878	0,39	-0,279	0,59	0,228	0,79	0,806	0,99	2,326

0,991	0,992	0,993	0,994	0,995	0,996	0,997	0,998	0,999
2,366	2,409	2,257	2,512	2,576	2,652	2,748	2,878	3,090

Б. Функция распределения стандартного нормального закона

$F(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{t^2}{2}}dt$ (таблица, обратная к предыдущей)

Для $x < 0$ следует пользоваться формулой: $F(-x) = 1 - F(x)$)

(например, $F(-2) = 1 - F(2) = 1 - 0,977 = 0,023$)

x	F(x)	x	F(x)	x	F(x)	x	F(x)
0,0	0,500	1	0,841	2,0	0,977	3,0	0,998
0,1	0,540	1,1	0,864	2,1	0,982	3,1	0,999
0,2	0,579	1,2	0,885	2,2	0,986	3,2	0,9993
0,3	0,618	1,3	0,903	2,3	0,989	3,3	0,9995
0,4	0,655	1,4	0,919	2,4	0,992	3,4	0,9997
0,5	0,691	1,5	0,933	2,5	0,994	3,5	0,9998
0,6	0,726	1,6	0,945	2,6	0,995	3,6	0,9998
0,7	0,758	1,7	0.955	2,7	0,996	3,7	0,9999
0,8	0,788	1,8	0,964	2,8	0,997	3,8	0,9999
0,9	0,816	1,9	0,971	2,9	0,998	3,9	0,9999

Таблица 4

Значения функции $\Phi(x) = \frac{2}{\sqrt{2\pi}} \cdot \int_0^x e^{-\frac{t^2}{2}} dt$

x	0	1	2	3	4	5	6	7	8	9
0,0	0,000	0,0080	0,0160	0,0238	0,0319	0,0399	0,0478	0,0558	0,0638	0,0717
0,1	0797,	0876	0955	1034	1113	1192	1271	1350	1428	1507
0,2	1585	1663	1741	1819	1897	1974	2051	2128	2205	2282
0,3	2358	2434	2510	2586	2661	2737	2812	2886	2960	3035
0,4	3108	3182	3255	3328	3401	3473	3545	3616	3688	3759
0,5	3829	3899	3969	4039	4108	4177	4245	4313	4381	4448
0,6	4515	4581	4647	4713	4778	4843	4907	4971	5035	5098
0,7	5161	5223	5285	5346	5407	5467	5527	5587	5646	5705
0,8	5763	5821	5878	5935	5991	6047	6102	6157	6211	6265
0,9	6319	6372	6424	6476	6528	6579	6629	6679	6729	6778
1,0	6827	6875	6923	6970	7017	7063	7109	7154	7199	7243
1,1	7287	7330	7373	7415	7457	7499	7540	7580	7620	7660
1,2	7699	7737	7775	7813	7850	7887	7923	7959	7994	8029
1,3	8064	8098	8132	8165	8198	8230	8262	8293	8324	8355
1,4	8385	8415	8444	8473	8501	8529	8557	8584	8611	8638
1,5	8664	8690	8715	8740	8764	8789	8812	8836	8859	8882
1,6	8904	8926	8948	8969	8990	9011	9031	9051	9070	9090
1,7	9109	9127	9146	9164	9181	9189	9216	9233	9249	9265
1,8	9281	9297	9312	9327	9342	9357	9371	9385	9399	9412
1,9	9426	9439	9451	9464	9476	9488	9500	9512	9523	9534
2,0	9545	9556	9566	9576	9586	9596	9606	9616	9625	9634
2,1	9643	9651	9660	9668	9676	9684	9692	9700	9707	9715
2,2	9722	9729	9736	9743	9749	9756	9762	9768	9774	9780
2,3	9786	9791	9797	9802	9807	9812	9817	9822	9827	9832
2,4	9836	9841	9845	9849	9853	9857	9861	9865	9869	9872
2,5	9876	9879	9883	9886	9889	9892	9895	9898	9901	9904
2,6	9907	9910	9912	9915	9917	9920	9922	0024	9926	9928
2,7	9931	9933	9935	9937	9938	9940	9942	9944	9946	9947
2,8	9949	9951	9952	9953	9955	9956	9958	9959	9960	9961
2,9	9963	9964	9965	9966	9967	9968	9969	9970	9971	9972

x	Φ(x)	x	Φ(x)	x	Φ(x)	x	
3,0	0,9973	3,4	0,9993	3,8	0,9999		
3,1	0,9981	3,5	0,9995	3,9	0,9999		
3,2	0,9986	3,6	0,9997	4,0	0,999936		
3,3	0,9990	3,7	0,9998	4,5	0,999994		

Таблица 5

Значения k_β нормального распределения

β	$\alpha=1-\beta$	$\tilde{\alpha}=\alpha/2$	k_β
0,9	0,1	0,05	1,65
0,95	0,05	0,025	1,96
0,98	0,02	0,01	2,3
0,99	0,01	0,005	2,58
0,9975	0,0025	0,00125	3,02

Принятые обозначения: α – ошибка, β – уровень доверия.

В таблице заданы k_β – решения уравнения $\beta = \frac{1}{\sqrt{2\pi}} \int_{-k_\beta}^{+k_\beta} e^{-\frac{x^2}{2}} dx$ (значения β в первом столбце, соответствующая ошибка $\alpha = 1-\beta$ во втором столбце, так что если задана ошибка α, то надо искать ее во втором столбце). Если нужно строить одностороннюю область, т.е. надо решать уравнение: $\tilde{\alpha} = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-k_{\tilde{\alpha}}} e^{-\frac{x^2}{2}} dx = \frac{1}{\sqrt{2\pi}} \int_{+k_{\tilde{\alpha}}}^{\infty} e^{-\frac{x^2}{2}} dx$, то $\tilde{\alpha} = 1-\beta$ надо искать в третьем столбце.

Таблица 6

Значения $t_{n;\beta}$ распределения Стьюдента

Число степеней свободы	Уровень значимости $\alpha = 1-\beta$ (двусторонняя критическая область). Заданы α/β			
n	0,10/0,90	0,05/0,95	0,02/0,98	0,01/0,99
1	6,31	12,7	31,82	63,7
2	2,92	4,30	6,97	9,92
3	2,35	3,18	4,54	5,84
4	2,13	2,78	3,75	4,6
5	2,01	2,57	3,37	4,03
6	1,94	2,45	3,14	3,71
7	1,89	2,36	3,00	3,5
8	1,86	2,31	2,9	3,36
9	1,83	2,26	2,82	3,25
10	1,81	2,23	2,76	3,17
11	1,80	2,220	2,72	3,11
12	1,78	2,18	2,68	3,05
13	1,77	2,16	2,65	3,01
14	1,76	2,14	2,62	2,98
15	1,75	2,13	2,60	2,95
16	1,75	2,12	2,58	2,92
17	1,74	2,11	2,57	2.90
18	1,73	2,10	2,55	2,88
19	1,73	2,09	2,54	2,86
20	1,73	2,09	2,53	2,85
30	1,70	2,04	2,46	2,75
40	1,68	2,02	2,42	2,70
60	1,67	2,00	2,39	2,66
120	1,66	1,98	2,36	2,62
∞	1,64	1,96	2,33	2,58
α	0,05/0,95	0,025/0,975	0,01/0,99	0,005/0,995
	Уровень значимости $\alpha = 1-\beta$ (односторонняя критическая область)			

Принятые обозначения: α – уровень значимости (ошибка), β – уровень доверия.

В таблице заданы $t_{n;\beta}$ – решения уравнения $\beta = \int_{-t_{n,\beta}}^{+t_{n,\beta}} f_n(x)dx$ (значения β в верхней строке). Если нужно решать уравнение $\alpha = \int_{-\infty}^{-t_{n,\alpha}} f_n(x)dx = \int_{+t_{n,\alpha}}^{\infty} f_n(x)dx$, то α надо искать в нижней строке.

Изд. лиц. ЛР № 071765 от 07.12.1998
Казахстанско-Российский университет
Объем 8,5 п.л. Тираж 754